Advanced Method of Synthesis of Semantic Kernel of E-content

Sergey Orekhov

National Technical University "Kharkiv Polytechnic Institute", Kyrpychova str. 2, Kharkiv, 61002, Ukraine

Abstract

The work describes an improved method of synthesis of the semantic kernel of e-content. This method is an integral part of the new technology of virtual promotion of goods and services. This technology is an alternative way to solve the problem of search engine optimization on the Internet. Its main components are the semantic kernel and the progress map. The previously obtained results of the execution of WEB projects of virtual promotion show that the correct choice of the semantic kernel gives an improvement of the main WEB metrics tenfold. Therefore, solving the problem of nuclear synthesis is an urgent problem. In this article, a synthesis method is proposed, which is based on a text model that includes aspect terms and collocations. The algorithm and results of ego application in a real WEB project for the virtual promotion of online services in the US market are also presented.

Keywords 1

Semantic kernel, virtual promotion, search engine optimization, bag of words

1. Introduction and related works

For the first time in works [1-2] the concept of semantic kernel was used to solve the problem of text classification. This task aims at automatic organization of text documents according to given categories. To solve it, each text document is presented in the form of a so-called "bag of words" (BOW). The BOW approach is simple and is used very often to describe the semantic kernel. Its main limitation is that it assumes independence between terms, since documents in the BOW model are represented by their terms, ignoring their position in the document, their semantic or syntactic relationships between other words.

The BOW model clearly does not take into account multi-word statements, breaking them into parts. In addition, it treats polysemous words (that is, words with multiple meanings) as a single entity. For example, the term "organ" can have the meaning of a body part when it appears in a context related to a biological structure, or the meaning of a musical instrument when it appears in a context related to music.

The work [2] states that each class of terms, that is, each "bag of terms" has two types of vocabulary: one is a "main" vocabulary closely related to the subject of this class, the other type is a "general" vocabulary, which may have a similar distribution in different classes. Thus, two documents from different classes can have many words in common and can be considered similar to the BOW representation.

To solve these problems, several methods have been used that use a measure of relatedness between terms in the areas of word meaning determination, text classification, and information retrieval. Semantic relatedness computation can be fundamentally divided into three categories, such as knowledge-based systems, statistical approaches, and hybrid methods that combine both ontology-based and statistical information.

Knowledge-based systems use a thesaurus or ontology to improve the representation of terms by taking advantage of the semantic relatedness between terms [3-6]. For example, in [4] the distance between words in WordNet is used to detect semantic similarity between English words. The study [4] uses superconcept declaration with different distance measures between words from WordNet,

MoMLeT+DS 2022: 4th International Workshop on Modern Machine Learning Technologies and Data Science, November, 25-26, 2022, Lviv-Leiden, Ukraine

EMAIL: sergey.v.orekhov@mail.com (S. Orekhov)
ORCID: 0000-0002-5040-5861 (S. Orekhov)

such as inverted path length, Wu-Palmer measure, Resnik measure, and Lin measure. The second type of computation of semantic relatedness between terms is corpus-based systems, in which some statistical analysis is performed based on the relations of terms in a set of training documents to reveal hidden similarities between them [7]. One of the well-known corpus systems is Latent Semantic Analysis (LSA) [8], which partially solves the synonymy problem. Finally, approaches in the last category are called hybrid because they combine information derived from both ontology and statistical corpus analysis [3].

Also in work [2] it was proposed to analyze the text corpus through the creation of a number of semantic cores, which are divided into groups: higher-order cores, iterative semantic cores, and lower-order cores. This classification makes it possible to improve the efficiency of document classification compared to traditional ones. Semantic kernels of the following types are distinguished: linear, polynomial and RBF kernels, when taking advantage of higher-order relationships between terms and documents. The basis for such a classification is the classic model of generating queries to the document database within the search server. This is a vector-space model (vector space model) [9]. Thus, vector-space model metrics are used to assess the degree of connectivity of terms in the "bag of terms".

As you know, the semantic core is a set of keywords formed on the basis of a given criterion. An example of such a criterion can be semantic relatedness between terms. But the approach proposed above cannot be used in the case of virtual promotion, because the semantic kernel to be formed includes a set of terms describing a given class of needs. Such a set of terms is connected not only by semantics, but more by the need of a potential client to buy a given product. That is, the connection between the terms is formed more due to market events, due to the laws of marketing.

Thus, in the technology of virtual promotion, it is necessary to form a semantic kernel as a set of keywords that are connected to each other both semantically and on the basis of market events, which are confirmed by the "4P" principle.

In virtual promotion, the semantic kernel plays two important roles.

The first area of kernel usage is messages in the promotion channel. The kernel describes the product being sold. It should have a semantic affinity with the class of need. That is, it is still necessary to have an additional core that describes the need that this product covers in the life of a potential buyer. There can be any number of such classes of needs. In addition, each kernel, according to marketing principles, should have keywords that emphasize its uniqueness for the buyer. Thus, on the one hand, the semantic kernel has semantic affinity with popular queries on the Internet, on the other hand, the kernel has unique elements that separate it from others. This allows you to separate one product from another in the mind of a potential buyer.

The work [11] proved the need to divide keywords from the kernel into classes, according to popular queries in search servers. Such classes are formed according to the main requests from the buyer: what needs are covered, where and when the product can be purchased. Such a structure is acceptable for the buyer, because it helps him quickly find the product and compare it with analogues.

The second is a means of managing the process of virtual promotion. By changing the kernel, it is possible to improve the results of the implementation of our technology, which was shown in works [11-13]. It was established that the gradual increase of the semantic kernel in the promotion channel leads to an increase in the value of the main WEB metric - the traffic of the WEB resource.

Given these prerequisites, let's build a mathematical model of semantic kernel based on the text model [11].

2. Mathematical model of semantic kernel

As input data we have a text document *D*, which can be written in any language and presented as HTML or plain text:

$$D = \{s_1, ..., s_n\}, \quad s_i = kw_i \cup gw_i \cup es_i, \tag{1}$$

where s_i, $i = \overline{1,n}$ is a sentence that ends with a comma or other end-of-sentence sign. Each sentence is a combination of three sets: a set of product description keywords kw_i, a set of sentence ending characters gw_i, and a set of keywords es_i that express the overall content.

Based on the approach [10], when the user's query should include the answer to three questions (what, where, when) to describe the product, formula (1) can be simplified.

$$D = \{s_1, ..., s_n\}, \quad s_i = kw_i \cup gw_i, \; s_i = \{w_{i1}, ..., w_{im}\}, \tag{2}$$

That is, you can ignore the set es_i. But for selecting keywords, you can use morphological homonymy exclusion methods to build semantic kernels from the first to at least the fifth order. Or use unigram or trigram and higher models.

This will make it possible to build semantic kernels that include more than two semantically related keywords. In addition, such a set of words (for example, three or more) guarantees a more detailed description of the product, which means that it strengthens the marketing component, that is, the connection of keywords from a marketing point of view. A typical trigram model looks like this:

$$w_{ij} = \arg\max P(w_{ij} | w_{ij-1}, w_{ij-2}) P(w_{ij} | w_{ij-1}, w_{ij+1}) P(w_{ij} | w_{ij+1}, w_{ij+2}) . \tag{3}$$

Formula (3) uses probability to link three key words into a single complex. In this case, the probability can be calculated as follows:

$$P(w_{ij} | w_{ij+1}) = \frac{F(w_{ij}, w_{ij+1})}{F(w_{ij})}, \tag{4}$$

where w_{ij}, w_{ij+1} keywords, $F(w_{ij}, w_{ij+1})$ is the frequency of occurrence of two words together, $F(w_{ij})$ is the total frequency of occurrence of the word w_{ij} w in the text document *D*.

In addition, it is advisable to also consider the parameters of the text document *D* itself. Among which the following should be noted:

$$S = \frac{P_n + P_p}{P_a + P_v}, \tag{5}$$

$$Q = \frac{P_a + P_{adv}}{P_n + P_v}, \tag{6}$$

$$A = \frac{P_v}{N}, \tag{7}$$

$$Di = \frac{P_v}{P_a + P_n + P_p}, \tag{8}$$

$$Z = \frac{P_C}{P_S}, \tag{9}$$

where *S* is the objectivity of the text document, *Q* is its quality, *A* is activity, *Di* – dynamism of the text document, *Z* – coherence, P_n – number of nouns, P_a – number of adjectives, P_v – number of verbs and verb forms (participle, adverb), P_p – number of pronouns, P_{adv} – number of adverbs, P_C – the number of prepositions and conjunctions, P_S – the number of independent sentences in the text, *N* – the number of words in the text document.

Why should homonyms be analyzed? Because probably each product has several situations (needs) when it should be used. Then, if there are several descriptions of the need, that is, classes of need in the given product, then the semantic kernel specifying the product may belong to one of these classes.

By changing the kernel, you can go virtually from one class of need to another. Thus, the analysis of homonyms guarantees the correct selection of classes of needs and their connection with semantic kernels.

Formulas (5)-(9) make it possible to analyze a text document for the purpose of identifying classes of needs as the first step on the way to building a semantic kernel.

In computational linguistics, syntactically correct word combinations that are stable in a statistical sense are usually called collocations. Most multi-word terms are collocations. The *MI* measure and its modifications are most often used to identify terms as collocations:

$$MI = \log_2 \frac{f(a,b)N}{f(a)f(b)}, \tag{10}$$

where *N* is the number of words in the text document. Function *f(a,b)* is the frequency of co-occurrence of words *a* and *b*, which evaluates the degree of dependence of the occurrence of two words in the corpus on each other. Functions *f(a), f(b)* are the frequencies of occurrence of words *a* and *b* separately from each other.

If the identified two-word collocations are considered as a whole, then with the help of the mentioned measures, longer word combinations (three-word, four-word, etc.) can be recognized in the text, which allows to extract long terms with an arbitrary syntactic structure using statistical criteria.

Thus, two mechanisms are established. The first is how to exclude or find homonyms. The second is how to evaluate collocations. Collocation is the ideal means of describing the semantic kernel. That is, our task is to identify collocations excluding homonyms.

Then, in advance, the task of forming the semantic core is formulated as follows: by evaluating (5)-(9) of the text document *D* for the presence of homonyms, determine collocations of key words using the given metric (10).

The construction of the method of forming the semantic kernel faces two problems. The first is to establish the semantic relatedness of keywords in the kernel. The second problem is that metrics (5)-(10) depend on the number of words in a text document. The fact is that modern marketing strategies aim to create descriptions of goods and services in an abbreviated style.

Therefore, the specified metrics may not accurately express the text attributes we need.

To solve the specified tasks, you can apply the mechanism of determining aspect keywords, because they most often create collocations and precisely for the definition of goods and services.

If both individual nouns and noun groups are extracted as aspect terms, it is necessary to use additional features to more accurately determine the length of the noun group. Most often, so-called contextual features are used, which estimate the frequency of occurrence of a word combination with the frequency of the context. Such signs allow you to determine the boundaries of the nominal group. For example, the so-called FLR measure is used:

$$FLR(a) = f(a)LR(a), \tag{11}$$

$$LR(a) = \sqrt{l(a)r(a)}, \tag{12}$$

where *f(a)* is the frequency of appearance of the aspect keyword *a*. Function *l(a)* is the number of different words to the left of *a*. Function *r(a)* is the number of different words to the right of *a*.

Next, noun groups with this measure greater than y are selected average for phrases. Thus, this measure primarily selects nouns that have a large variety of words at their boundaries, indicating that the analyzed term a is not a fragment of a longer phrase.

Another criterion aimed at the same goal is the well-known *C-value* metric [11], which reduces the weight of a given word or phrase if it is part of a longer frequency phrase. Thus, it is assumed that this longer phrase can be considered as a candidate aspect, and the current one represents its fragment. Such a sign for selecting aspects is used in the work.

Then finally, the task of forming a semantic kernel can be defined as follows: given a text document D consisting of a number of sentences, select a set of aspectual collocations of at least the third order to ensure a rational level of semantic and marketing relatedness.

3. Proposed method

First, let's describe the method of formation verbally. As input information, we have a text document *D* given by formula (1). But the set es_i is redundant, so it must first be removed to obtain a document of the form (2). Then we introduce the normalization and lemmatization operation: $NL: D \to D_{nl}$. Exactly $D_{nl} = KW \cup GW$. The set of keywords of general content cannot be excluded, because these words play the role of a link between the main aspect terms.

Next, the operation of determining aspect terms and the operation of determining collocations should be applied to the set $D_{nl} = KW \cup GW$: $A: D_{nl} \to D_{nl}^{a}$ and $K: D_{nl}^{a} \to D_{nl}^{ak}$. As a result, a set is formed $D_{nl}^{ak} = KW_{nl}^{ak} \cup GW_{nl}^{ak}$. This set will be the source of semantic kernels.

However, the presence of a set of keywords that can be part of the semantic kernel is not enough in our case. The reason is that the formation of this set was performed only taking into account the semantics and properties of the document *D*. Even if the text corpus $TDC = \{D_1,...,D_C\}$ is considered, where $D \in TDC$. We must also consider the other side – WEB statistics of the use of plural keywords from D_{nl}^{ak}. Therefore, we add the following operation, the purpose of which is to form an estimate of each keyword and each word combination that can be formed on the basis of the set D_{nl}^{ak}. We introduce the evaluation metric based on VEB statistics:

$$F_{web} : D_{nl}^{ak} \to M_{web},\ M_{web} \in R,\ M_{web} \geq 0.$$

As a result, pairs $SK = \{(D_{nl}^{ak}, M_{web})_1,...,(D_{nl}^{ak}, M_{web})_C\}$ are formed for each document of the *TDC* text corpus. We will assume that keywords with the maximum value of the metric M_{web} fall into the semantic kernel:

$$sk = \max_{w \in TDC} M_{web}(w),$$

where *w* is a phrase that includes at least two keywords.

Thus, to form a kernel, the following operations must be performed in sequence: normalization, lemmatization, aspectization, collocation, and evaluation. In addition, it was shown above that it is not enough to have only a semantic kernel. It is necessary to first form classes of needs, that is, sets of keywords that describe unique needs that are covered by these other products. These are components of a marketing strategy, where it is described that a given product or service can be used for an urgent need of a given group of customers.

Then, first of all, the text corpus should be analyzed for identifying classes of needs. For this, metrics (5)-(9) should be used. Accordingly, our method should be supplemented with one more step - finding a class of needs.

The work suggests that one document contains a description of the product for sale, or a description of the need that this product covers. It is also possible with high probability that one document contains both descriptions at once. Then we will assume that the text corpus contains 20% of documents expressing the main idea about the need and the product. In order to find this set of documents, it is necessary to rank them. Then it is suggested to first calculate for each document the value of metrics (5)-(9). A document with the maximum value contains keywords that describe as fully as possible one of the classes of need and the product that covers it.

As a result, two tasks are formed. The first task is to organize a set of documents in order to select a list of documents where keywords should be searched for the formation of semantic cores. The order determines the priority according to which the classes in our study are organized. That is, the single document with the best number is the document that contains a separate requirement class.

The second task involves searching for documents where there is no description of need classes, but there is a description of semantic kernels that belong to some need class. That is, this set of documents, as it were, functions to confirm the existence of a class of need and a product that covers it.

This structure of tasks is based on the classical method of searching Page Rank [14]. The main principle of implementing this method is that the WEB resource that has the maximum number of links from other WEB resources gets the maximum result. Based on the Pareto principle, we will select 20% of the documents with the maximum values of the document quality assessment metrics (5)-(9), because they contain 80% of all the semantic connections and marketing knowledge we need [15].

4. Proposed algorithm

The paper proposes a method for forming the semantic kernel, which includes two cycles. Consider the first loop and its verbal algorithm.

Stage 1. We take the *i*-th document of the *TDC* corpus for processing. We calculate the metrics (5)-(9).

Stage 2. We build table 1, which accumulates the list of documents and values of metrics (5)-(9). We perform these two stages until table 1 is completely filled with all the documents that are available in the text corpus.

Table 1

Search for documents describing classes of needs

Document Number	*S*	*Q*	*A*	*Di*	*Sum*

Stage 3. We calculate the sum of these metrics. We sort them all documents of the text corpus according to the maximum value of the "Sum" column. As a result, we choose the first values of this column, approximately 20% of the total number of documents. Thus, the first documents in terms of the value of the last column express the classes of need and the description of the product that belongs to them covers. These documents will form a set of TDC''. It can be used as input information for the second cycle of our method to build a set of keywords from which the semantic core of e-content will be formed.

Consider the second cycle. We will assume that the input information is a set of documents $D'' \in TDC''$. The following steps are suggested.

Stage 1 (normalization). We take the *i*-th document of the corpus for processing $D_i'' \in TDC''$. From all the sentences of the document D_i'', we select three groups of keywords $s_{ij}'' = kw_{ij}'' \cup gw_{ij}'' \cup es_{ij}''$. Next, we exclude the plural of words es_{ij}'' completely. As a result each corpus document includes two groups of keywords: aspect terms and general content words ($s_{ij}'' = kw_{ij}'' \cup gw_{ij}''$, $s_{ij}'' \in D_i''$).

Stage 2 (search for collocations). We build table 2, which allows us to identify the list of candidates for collocations of various orders. In table 2, we enter the words $w'' \in kw_{ij}'' \cup gw_{ij}''$ and determine the meaning of metric (10).

As you can see, Table 2 is a matrix that allows you to identify two-word collocations. But for our research it is also important to know about the existence of three- and four-word collocations. Therefore, at this stage, having obtained two-word collocations, it is necessary to rebuild Table 2 in order to iteratively analyze more complex combinations of keywords. This process continues until the values of the metric (10) remain unchanged, that is, all complex collocations are obtained.

Table 2

Candidates for collocations

Metric *MI*	*Candidate 1*	*Candidate 2*	*Candidate 3*	…	*Candidate M*
Candidate 1					
…					
Candidate M					

We will denote the set of collocations as follows: $D_i'' = \{col_i^2\} \cup \{col_i^3\} \cup ... \cup \{col_i^R\}$, where col_i^2 – collocations of the second order, col_i^3 – collocations of the third order and so on.

But it is empirically shown that for virtual promotion it is enough that $R \le 5$. This is due to the fact that queries in search engines of more than five words have a very low probability [16].

Stage 3 (search for aspect terms). Again, we consider each document in the document *D* the keywords $w'' \in kw_{ij}'' \cup gw_{ij}''$. We calculate the metrics (11)-(12) for each candidate in Table 3.

The largest values of metrics (11)-(12) allow, again using the Pareto principle, determine about 20% of aspect terms from their total number in the document $D_i'' = \{w''\}_i$. Next, if we compare the data of tables 2 and 3, we get a list of collocations, where aspect terms are present.

These collocations are the first candidates for the semantic kernel of e-content. That is, we get an intermediate result: $D_i'' = \{col_i^{a2}\} \cup \{col_i^{a3}\} \cup ... \cup \{col_i^{aR}\}$.

Table 3
Candidates for aspect terms

No	*Candidate (aspect term a)*	*f(a)*	*FLR(a)*	*LR(a)*	*Priority*

Stage 4 (adding aspect terms). At the same time, having only collocations in the semantic kernel is inefficient, so it is necessary to add to the set of candidates AT_i some aspect terms that are not included in the final set D_i'' for each document: $D_i'' = \{col_i^{a2}\} \cup \{col_i^{a3}\} \cup ... \cup \{col_i^{aR}\} \cup \{AT_i\}$.

Having the semantic kernels of the e-content of a given WEB resource, it is possible activate the process of virtual promotion on the Internet. Because, as was shown in [17], the semantic kernel is the message and driving impulse in our virtual promotion.

The action diagram that describes the proposed algorithm is presented in Figure 1.

Consider an example calculated according to the proposed algorithm based on the results of the WEB project.

5. Results

Let's consider the initial conditions that existed at the start of the WEB project for the American market of WEB services. This project was started to meet the need of users to build a psychological portrait of an individual online. The start of this WEB project took place when classical methods of search engine optimization gave no effect.

The desired effect was primarily perceived as an increase in the number of users of this Web service. This goal can be achieved by increasing the number of visits to this WEB resource. Therefore, the synthesis of semantic kernel was offered to the owner of this WEB resource in order to increase the value of the WEB metric - traffic.

At the time of the start of the test WEB project (Figure 2), the value of the metric was minimal. Therefore, it was proposed to use a new approach (kernel synthesis) in order to increase traffic with its help.

Table 4 shows a fragment of the e-content of this site at the time of the start of synthesis. This WEB resource included a list of 13 documents. Only nine of them contain text e-content about the market, need or product. At the first stage, the first cycle of the semantic kernel synthesis algorithm was performed, which is specified in section 4. Table 5 contains the results of processing these nine documents. According to the algorithm, we will select two documents to create a set of keywords that will be candidates for the semantic kernel. They are shown also in table 5.

Next, we perform the second cycle of the synthesis algorithm. To do this, we remove all words except plural nouns, adjectives, adverbs and verbs. We are normalizing these words. But first, we choose the document with the maximum value of the document quality metrics - the eighth line of table 5. Test e-content from this page is presented in table 4. Figure 3 presents the keywords and calculation of the MI measure for this e-content.

The data in Figure 3 demonstrate the fact that due to the small number of keywords in e-content, it is not possible to detect collocations. Therefore, we proceed to the next step of our algorithm, namely, the analysis of aspect terms.

The results of the search for aspect keywords are based on the calculation of metrics, which are shown in Figure 4. The following sequence of aspect keywords was revealed: "CelestialTiming, easy, accurate, tool, bring, user, site, natural, rhythm, providing, goal, universe, effective, collaboration". Thus, our algorithm forms a set of candidates for the semantic kernel of e-content.

Within this test project, the following strategy for launching the semantic kernel was proposed. At the first stage, due to the fact that this web resource was a startup, it is necessary to implement a semantic core from only one aspect keyword - CelestialTiming. Because potential users of this service do not have any information about this startup. At the second stage, it is necessary to expand the semantic kernel with the following words from the found set.

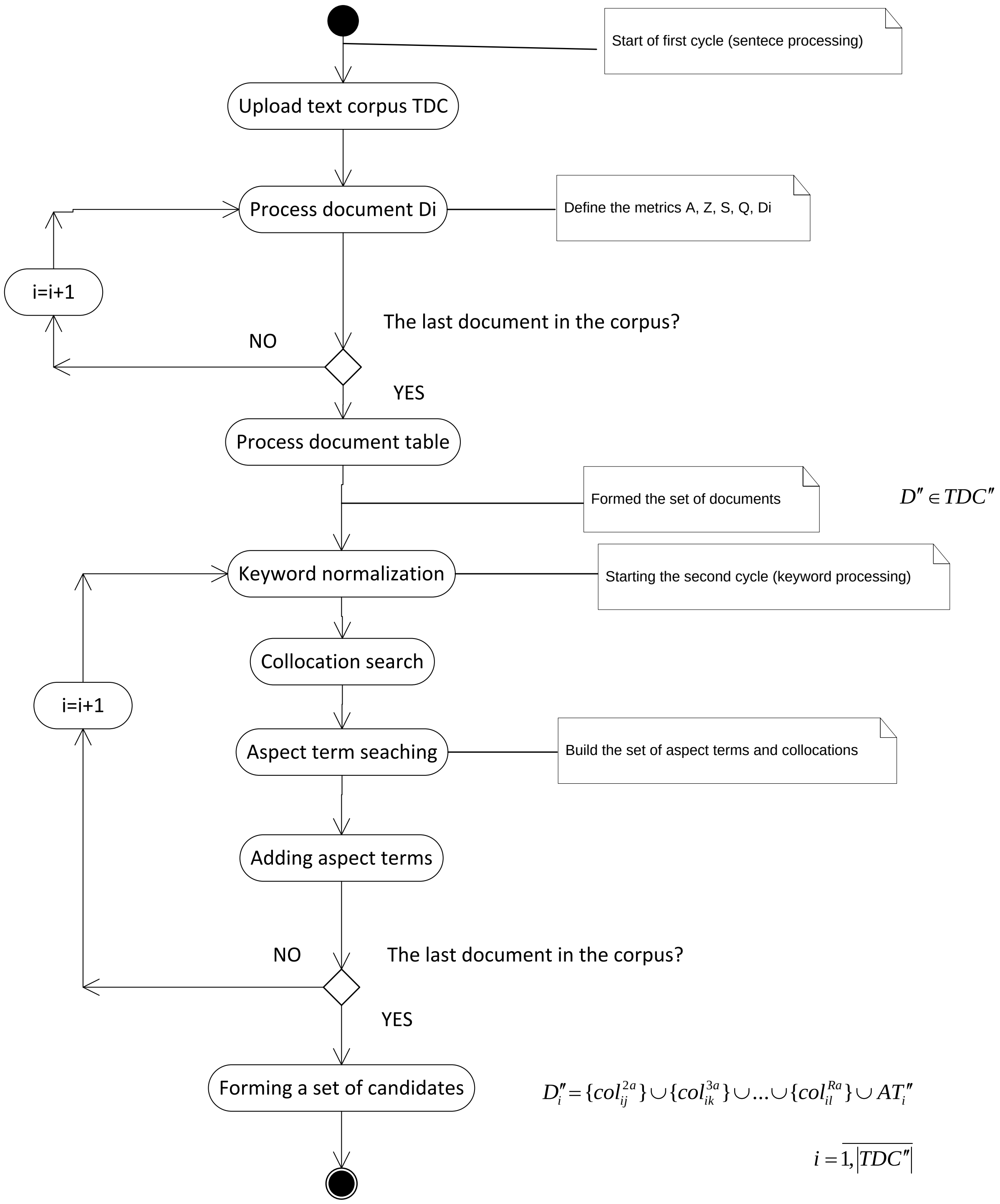

Start of first cycle (sentece processing)
Upload text corpus TDC
Process document Di
Define the metrics A, Z, S, Q, Di
i=i+1
The last document in the corpus?
NO
YES
Process document table
Formed the set of documents
$D'' \in TDC''$
Keyword normalization
Starting the second cycle (keyword processing)
i=i+1
Collocation search
Aspect term seaching
Build the set of aspect terms and collocations
Adding aspect terms
NO
The last document in the corpus?
YES
Forming a set of candidates
$D_i'' = \{col_{ij}^{2a}\} \cup \{col_{ik}^{3a}\} \cup \ldots \cup \{col_{il}^{Ra}\} \cup AT_i''$
$i = \overline{1, |TDC''|}$

Figure 1: Activity diagram

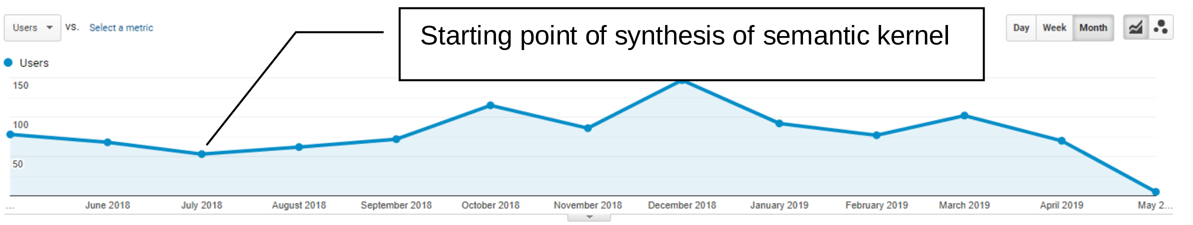


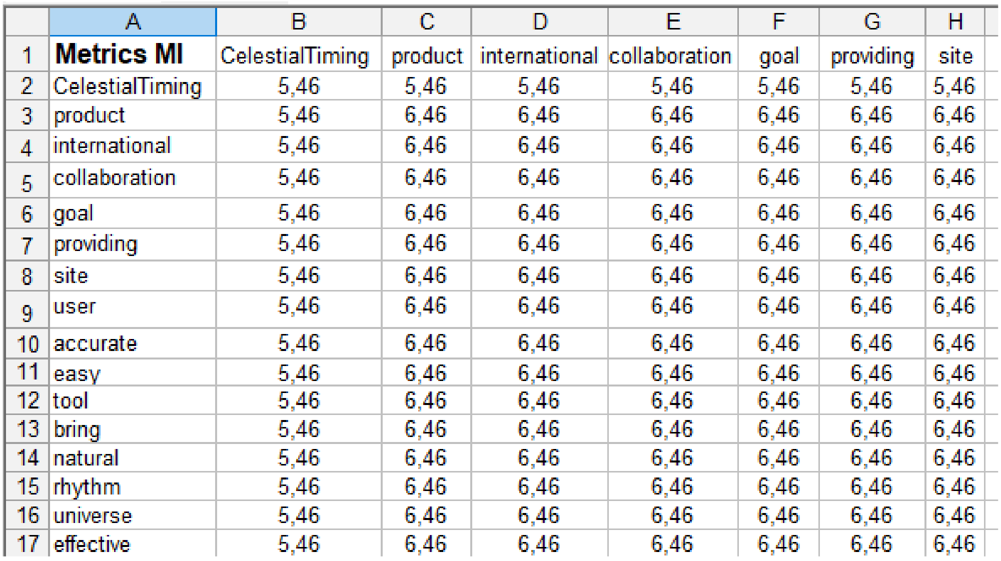
Figure 2: Traffic statistics of test WEB project

Table 4
A fragment of the e-content of the first test WEB resource

No	*E-content*	*Comment*
1	CelestialTiming is the product of international collaboration with a goal of providing site users with an accurate and easy tool to bring them in touch with their natural rhythms of the universe and enable them to make more effective personal and business decisions. Members of the CelestialTiming team have extensive experience and advanced degrees in the fields of physics, engineering, computer science, psychology, and education. They have developed professional psychological and astrological software, taught in colleges and universities, and published books and articles on a variety of topics…	Description of the main product

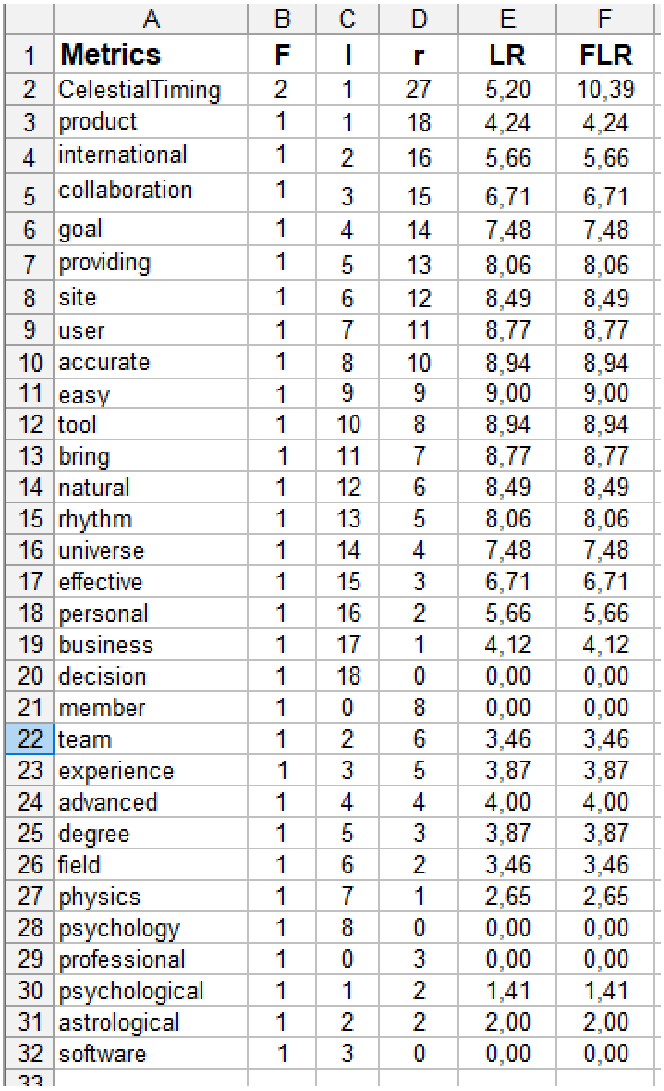
Table 5
Search for documents describing classes of needs

No	*Web page*	*S*	*Q*	*A*	*Di*	*Z*	*Sum*
1	Home	1.336	0.265	0.127	0.347	0.417	2.491
2	celebrity	1.196	0.339	0.138	0.415	2.125	4.213
3	myself	0.778	0.455	0.152	0.455	2.0	3.838
4	another	0.636	0.5	0.182	0.5	2.0	3.818
5	saved	0.83	0.451	0.152	0.403	1.609	3.445
6	applications	1.071	0.287	0.130	0.477	2.764	4.728
7	**free**	**1.263**	**0.407**	**0.125**	**0.303**	**3.4**	**5.499**
8	**contact**	**1.5**	**0.469**	**0.057**	**0.111**	**7.667**	**9.803**
9	account	0.667	0.455	0.15	0.471	2.25	3.991



	A	B	C	D	E	F	G	H
1	**Metrics MI**	CelestialTiming	product	international	collaboration	goal	providing	site
2	CelestialTiming	5,46	5,46	5,46	5,46	5,46	5,46	5,46
3	product	5,46	6,46	6,46	6,46	6,46	6,46	6,46
4	international	5,46	6,46	6,46	6,46	6,46	6,46	6,46
5	collaboration	5,46	6,46	6,46	6,46	6,46	6,46	6,46
6	goal	5,46	6,46	6,46	6,46	6,46	6,46	6,46
7	providing	5,46	6,46	6,46	6,46	6,46	6,46	6,46
8	site	5,46	6,46	6,46	6,46	6,46	6,46	6,46
9	user	5,46	6,46	6,46	6,46	6,46	6,46	6,46
10	accurate	5,46	6,46	6,46	6,46	6,46	6,46	6,46
11	easy	5,46	6,46	6,46	6,46	6,46	6,46	6,46
12	tool	5,46	6,46	6,46	6,46	6,46	6,46	6,46
13	bring	5,46	6,46	6,46	6,46	6,46	6,46	6,46
14	natural	5,46	6,46	6,46	6,46	6,46	6,46	6,46
15	rhythm	5,46	6,46	6,46	6,46	6,46	6,46	6,46
16	universe	5,46	6,46	6,46	6,46	6,46	6,46	6,46
17	effective	5,46	6,46	6,46	6,46	6,46	6,46	6,46

Figure 3: A fragment of the table with calculations of the value of the MI metric



	A	B	C	D	E	F
1	**Metrics**	**F**	**I**	**r**	**LR**	**FLR**
2	CelestialTiming	2	1	27	5,20	10,39
3	product	1	1	18	4,24	4,24
4	international	1	2	16	5,66	5,66
5	collaboration	1	3	15	6,71	6,71
6	goal	1	4	14	7,48	7,48
7	providing	1	5	13	8,06	8,06
8	site	1	6	12	8,49	8,49
9	user	1	7	11	8,77	8,77
10	accurate	1	8	10	8,94	8,94
11	easy	1	9	9	9,00	9,00
12	tool	1	10	8	8,94	8,94
13	bring	1	11	7	8,77	8,77
14	natural	1	12	6	8,49	8,49
15	rhythm	1	13	5	8,06	8,06
16	universe	1	14	4	7,48	7,48
17	effective	1	15	3	6,71	6,71
18	personal	1	16	2	5,66	5,66
19	business	1	17	1	4,12	4,12
20	decision	1	18	0	0,00	0,00
21	member	1	0	8	0,00	0,00
22	team	1	2	6	3,46	3,46
23	experience	1	3	5	3,87	3,87
24	advanced	1	4	4	4,00	4,00
25	degree	1	5	3	3,87	3,87
26	field	1	6	2	3,46	3,46
27	physics	1	7	1	2,65	2,65
28	psychology	1	8	0	0,00	0,00
29	professional	1	0	3	0,00	0,00
30	psychological	1	1	2	1,41	1,41
31	astrological	1	2	2	2,00	2,00
32	software	1	3	0	0,00	0,00

Figure 4: Calculating the value of metrics for finding candidates for facet keywords

The results obtained (Figure 2) show that the introduction of a new semantic kernel, which symbolizes the brand of our online service, has led to an increase in traffic. That is, new users who were interested in learning about a new online service came to the WEB site. However, this effect was short-lived. This is also understandable, since the effect of aging of the semantic kernel was established in the works [11-12], which in this case manifested itself.

6. Summary

The paper demonstrates a new improved approach to the synthesis of the semantic kernel of e-content. This approach has been tested on a real WEB project in the US online services market. The article shows the effect of the implementation of the synthesis algorithm, which was positive for a real WEB project.

Thus, we can say that the following new results have been obtained:

- the task of forming the semantic kernel of e-content received further theoretical and methodological development;
- for the first time, it is proposed to apply metrics for evaluating texts for evaluating e-content keywords and forming a semantic kernel, taking into account the entire text corpus of a given WEB resource;
- the method of forming the semantic kernel of e-content, which includes two cycles, was further developed. The first is for establishing a list of documents for processing, and the second is for forming candidates (key phrases and words) for the semantic kernel.

In the future, it is planned to implement this method as a separate software component on the platform Node JS.

7. References

[1] Altinel B., Ganiz M. C., Diri B. A simple semantic kernel approach for SVM using higher-order paths. // IEEE International Symposium on Innovations in Intelligent Systems and Applications (INISTA) Proceedings. – 2014. – P. 431-435.

[2] Altınel B., Ganiz M., Diri B. A corpus-based semantic kernel for text classification by using meaning values of terms. Engineering Applications of Artificial Intelligence. 2015. – Vol. 43. – P. 54-66.

[3] Nasir J.A., Varlamis I., Karim A., Tsatsaronis G. Semantic smoothing for text clustering. // Knowledge-Based Systems. – 2013. – Volume 54. – P. 216-229.

[4] Budanitsky, A., Hirst, G. Evaluating WordNet-based measures of lexical semantic relatedness. // J. Computational Linguistics. – 2006. – Volume 32(1). – P. 13–47.

[5] Bloehdorn, S., Basili, R., Cammisa, M., Moschitti, A. Semantic kernels for text classification based on topological measures of feature similarity. // Proceedings of the Sixth International Conference on Data Mining (ICDM). – 2006. – P. 808–812.

[6] Luo Q., Chen E., Xiong H. A semantic term weighting scheme for text categorization. // Expert Systems with Applications. – 2011. – Volume 38(10). – P. 12708-12716.

[7] Zhang, Z., Gentile, A.L., Ciravegna, F. Recent advances in methods of lexical semantic relatedness–a survey. // Natural Language Engineering. – 2012. – Volume 1(1). – P. 1–69.

[8] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., Harshman, R. Indexing by latent semantic analysis. // Journal of the American society for information science. – 1990. – Volume 41(6). – P. 391-407.

[9] Salton G., Wong A., Yang C. A vector space model for automatic indexing. // Computer science. – 1975. – Volume 18. – P. 613-620.

[10] Orekhov S., Malyhon H., Liutenko I., Goncharenko T. Using Internet News Flows as Marketing Data Component. // CEUR-WS, 2020. – Volume 2604. – P. 358-373.

[11] Orekhov S., Malyhon H., Goncharenko T. Mathematical Model of Semantic Kernel of WEB site. // CEUR-WS, 2021. – Vol. 2917. – pp. 273-282.

[12] Orekhov S., Malyhon H., Stratienko N., Goncharenko T. Software Development for Semantic Kernel Forming. // CEUR-WS, 2021. – Vol. 2870. – P. 1312–1322.

[13] Godlevsky M., Orekhov S., Orekhova E. Theoretical Fundamentals of Search Engine Optimization Based on Machine Learning. // CEUR-WS, 2017. – № 1844. – P. 23-32.

[14] Dode A., Hasani S. PageRank Algorithm. // Journal of Computer Engineering. – 2017. – Volume 19, Issue 1. – P. 1-7.

[15] Koch R. The 80/20 Principle. The Secret of Achieving More with Less. – Great Britain: Nicholas Brealey Publishing Limited, 1998. – 313 p.

[16] Rowley J. Understanding digital content marketing // Journal of Marketing Management. – 2008. – T. 24. – Volume. 5-6. – P. 517–540.

[17] Orekhov S., Kopp A., Orlovskyi D. Map of Virtual Promotion of a Product. // Advances in Intelligent Systems, Computer Science and Digital Economics III. Switzerland: Springer, 2022. – pp. 1-11.