Mathematical Model of Semantic Kernel of WEB site

Sergey Orekhov, Henadii Malyhon and Tetiana Goncharenko

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

Abstract

Our latest research from search engine optimization projects shows the effect of the semantic kernel of a website. It is a unique set of keywords that depends on time as well as the current state of the search engine vector space model. Therefore, the problem of mathematical modeling of changes in the semantic kernel of a website within a period of a month - a year is becoming urgent. The kernel is formed from three clusters of keywords: product (service), geography (location) and time (duration). The article proposes a model for representing the semantic kernel of a website. This view is supported by a description of the kernel based on the semantic web, followed by its presentation in the form of a resource description framework (RDF) schema. An algorithm for forming a kernel based on hierarchical clustering (agglomerative nesting) is also considered.

Keywords 1

Semantic kernel, search engine optimization, clustering, degree of proximity

1. Introduction

Since 2009, our team has completed more than thirty projects in the field of search engine optimization. The projects were in various fields: pharmacy, marketing research, jewelry production, construction materials, cosmetics, wood products, furniture production, and automobile spare parts. All the projects were united by one circumstance, namely: they are all related to the field of e-commerce. One way or the other, but the main goal of the project was set as an increase in the volume of sales of goods or services via the Internet. Some of the projects were successful, and some were clearly with negative results. Analyzing the results obtained, we came to the following conclusion.

There is an interesting effect of the so-called learning of search engine [1]. The fact is that search engine optimization is the process of training a search engine to respond to given user requests by showing our website in the first places in the list of answers. For this purpose, according to the space model vector, texts from our website that describe a product or service must have the maximum number of external links (large citation index). However, such links can be divided into two groups: black (formally created for the citation index) and white (created by real users who, for example, have already tested this product or service). In our case, white links are of the greatest interest, but such links are generated based on semantics.

Let us ask ourselves a question: how does an ordinary user form such a link? That is right, he learns in advance by using a product or service and then he writes a review for such use. In other words, feedback is generated. In this way, the search engine responds to a user feedback on our website, or rather, he reacts to texts and images. This reaction is shaped by us, however, this is an indirect reaction. At the first stage, we recognize texts from websites. Analyzing these texts, we either agree with them or not. Our consent is expressed in a positive response to the website. The process of such analysis is clearly short-lived. There are a lot of texts and websites, so we react to so-called annotations, that is, short sets of keywords that, in our opinion, are as complete as possible, but briefly describe the meaning of the text on the website. We will call such short descriptions of texts – semantic kernels.

© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

MoMLeT+DS 2021: 3rd International Workshop on Modern Machine Learning Technologies and Data Science, June 5, 2021, Lviv-Shatsk, Ukraine

EMAIL: sergey.v.orekhov@gmail.com (S. Orekhov); gmalygon@gmail.com (H. Malyhon); tatianagoncharenko1806@gmail.com (T. Goncharenko)

ORCID: 0000-0002-5040-5861 (S. Orekhov); 0000-0001-5448-2488 (H. Malyhon); 0000-0001-6630-307X (T. Goncharenko)

CEUR Workshop Proceedings (CEUR-WS.org)

Thus, there is an urgent problem of mathematical modeling of the semantic kernel. It is also necessary to describe mathematically the algorithm for its formation from the text of the website.

2. Problem statement

The simplest approach to considering the semantic kernel is to represent the source text as a bag of terms [2] with its subsequent clustering. As it was established [3], we should distinguish three clusters of terms: geography, product (service) and time.

There are three main clustering approaches: probabilistic (statistical), graph and hierarchical algorithms [4-5].

Graph clustering algorithms represent the initial sample in the form of a graph, where the nodes of the graph are the sample objects themselves, and the edges are pairwise distances between the sample objects. The main advantage of graph clustering algorithms is clarity and ease of implementation, as well as the ability to make various improvements using simple geometric assumptions. How are the edges of graphs constructed? Moreover, what distinguishes one method from another?

Most graph algorithms work is based on a hypothesis about the number of clusters that should be obtained in the end. For example, Minimum Spanning Tree algorithm assumes that each new sample point is sequentially connected to the nearest one, which is already connected to others. Then k-1 longest edges are removed at the end of the algorithm execution. The parameter k in this case is a hypothesis about the number of clusters. Since, in principle, it is impossible to know how many event clusters can occur a day, the advancement of any hypotheses about their number will be unacceptable and will only distort the clustering results [6].

Probabilistic (statistical) algorithms assume that each object is a random variable. Then the cluster will obey some distribution law. However, since the form of the distribution law is unknown in advance, then for the operation of the algorithm it is necessary to put forward a hypothesis about the applicability of a certain distribution law. A hypothesis that is incorrectly put forward can significantly distort the clustering results.

Therefore, probabilistic algorithms are also an ineffective approach to solving the problem of clustering terms. Because, firstly, it is impossible to determine the distribution laws of terms, and secondly, the number of terms is finite and small enough in comparison with the number of terms that can be used to describe a given triad (product, time and location). Thirdly, the functioning of probabilistic algorithms also often requires an assumption about the number of clusters, the initial centers of clusters, which leads to a significant qualitative difference in the results of the algorithm.

Hierarchical (taxonomic) clustering algorithms form not one sample partitioning into disjoint clusters, but a system of nested partitions. The result of such an algorithm is presented in the form of a dendrogram (Figure 1).



Figure 1: Dendrogram of semantic kernel

Among the algorithms for hierarchical clustering, there are two main types, depending on the logic of building clusters: divisional and agglomerative [6]. Divisional or top-down algorithms split the

original sample into smaller and smaller clusters. Agglomerative or bottom-up algorithms are those in which objects are combined into larger and larger clusters.

In the study, having evaluated the advantages and disadvantages of different approaches, the agglomerative hierarchical algorithms was proposed to apply.

Let at the initial moment of the execution of the algorithm each object (term) be considered a separate cluster. Let us start the merging process, where at each iteration a new cluster $A_i \cup A_j$ is formed instead of a pair of the closest clusters A_i and A_j . The quality of the algorithm is:

- in the method of determining the distance between A_i and A_j ;
- in the method of determining the distance between the new cluster $A_i \bigcup A_j$ formed in the

previous step and some element A_f to be united.

At the same time, do not forget about the so-called dubbing situation, which is in the fact that the same terms, but from the perspective of different users, can describe the same triad: product, time and geography.

In addition, when describing a triad, the user uses the concept of a plot, that is, a certain scenario within which he applies a given product or service.

Consequently, when constructing an algorithm for identifying the semantic core, it is necessary to take into account the moment of terms duplication, as well as a possible storyline in which these terms take part.

To automatically highlight duplicate and plot terms, it is necessary to formulate the corresponding metric criteria for the proximity of terms.

We will formulate a vector interpretation of an object (term) based on data obtained in the process of processing web content and introduce metric criteria for the proximity of two terms and an algorithm for combining terms into one of the clusters: geography, goods and time.

Each term is assigned to one of three categories, reflecting the type of term. Let us designate a set of terms that form dictionaries of three categories, $V = \{v_j\}$, $j = \overline{1,J}$ where v_j is a term in the dictionary. For a set of three categories, we denote the corresponding dictionaries $V^k = \{v_j^k\}$, $k = \overline{1,3}$. At the same time, the same terms cannot form different dictionaries of categories $V^k \cap V^r = 0$; $k \neq r$; $r, k \in (1,3)$, which make it possible to unambiguously distinguish a category exclusively on dictionaries of terms. For example, $V^P = \{wood, metal, auto, ...\}$, $V^T = \{month, day, week, ...\}$ and $V^G = \{latitude, longitude, ...\}$.

The required identification accuracy is achieved due to the fact that the number of categories is strictly limited. Terms that fall into one category are unlikely to fall into any other.

Let the semantic kernel of web content be based on these dictionaries. That is, the kernel will include only terms from these three clusters (dictionaries). Thus, our task is to build a vector of semantic kernel by clustering existing terms from web content.

An analysis based on this approach allows, in a primary approximation, to evaluate the kernel described on the website, however, the allocation of a unique core is only possible with the complete processing of a group of websites of a given topic.

The problem of processing the flow of websites is that the most of the data is non-metric, and therefore it is necessary to switch from a non-metric representation of terms to a metric one, in which it is possible to describe a universal criterion for the proximity of two terms in one category.

In our approach, the processing of web content involves the formation of three vectors: a vector of terms $\dot{\vec{n}}$, reflecting information about: what product, where it is located and at what time; vector $\ddot{\vec{n}}$ is recovered on the basis; vector of terms $\ddot{\vec{n}}$ that specifies the vector $\dot{\vec{n}}$. Thus, these vectors describe a semantic kernel.

A vector $\dot{\vec{n}}$ always has a fixed structure: $\dot{\vec{n}} = (d, p, G, L, M)$, where d – contains the date and time (when), $p \in P$ – product (what), $g \in G$ – product geography (where), $l \in L$ – many sources of web content, $m \in M$ – many related products (services) mentioned in web content (with by whom).

All sets P, G, L, M contain terms found in the web content of a given web site. Each element of such a set takes on the value zero if the term is absent in the web content and the value one if it is present. Set *d* contains the date and time of web content creation.

The vector \vec{n} describes a variant of the semantic core. At the same time, the vector \vec{n} contains the terms synonyms that possibly exist in the web content. In addition, we need a third vector \vec{n} , which contains possible terms that are service terms to describe the main meaning of web content.

Then we will assume that the web content of the web resource contains several semantic cores, possibly close in meaning. Our task is to collect several cores from the terms of a bag of words by clustering and check them for dubbing. In addition, as a rule, the content management system of a web resource can contain several versions of the kernels, possibly also duplicating each other. In addition, the search server in its database contains several versions of the semantic cores of this web resource. And the third place, where a variant of the semantic core of a web resource is possibly contained, is a social network, or rather an account in a social network or marketplace. Thus, by clustering, we form a set of nuclei and it is necessary to establish whether they are close and by how much. This affinity is also important as it echoes the idea of linking to a document in a search engine.

Therefore, we need a metric indicating the similarity or duplication of cores both on the web resource and on the Internet as a whole. The first parameter for evaluation is probably the date and time of the publication of web content. Consider a metric based on this assumption.

To combine two kernels – duplicates in one dictionary, the primary condition is the coincidence of the categories of terms. The date and time of the term appearance in web content should differ within the threshold value $d_{treshold}$ reflecting the dynamics of the market. The degree of proximity for the rest of the values is carried out according to the formula (1):

$$F_1(\dot{\vec{n}}', \dot{\vec{n}}'') = (1 + \sum_{j=0}^J (p_j' - p_j'')^2)(1 + \sum_{j=0}^J (m_j' - m_j'')^2)(1 + \sum_{j=0}^J (l_j' - l_j'')^2)(1 + \sum_{j=0}^J (g_j' - g_j'')^2) \to \min , (1)$$

where \vec{M}', \vec{M}'' – vectors described by coordinates $M' \cup M''$. Each element has a value $m_j = (0,1)$ if a term has been possessed to a vector $\dot{\vec{n}}$;

 \vec{L}', \vec{L}'' – vectors of web content sources respectively are taken from two kernel variants;

 \vec{G}', \vec{G}'' – vectors of geography, while the values of the geography of the terms should be reduced to the total dimension.

Obviously, in most cases the vectors differ and complete coincidence, when the criterion equal to one, is not observed. This problem can be solved either by calculating expert estimates of the thresholds for combining terms, or by forming the threshold value in an analytical way.

The following analytical method for calculating the error coefficient α is proposed, based on the reconstructed vector \vec{n} of other terms. The set of such terms should include those that are included in the web content, but at the same time do not reflect the very description of the product or service directly, but only clarify its description, in particular, indicate secondary signs, characteristics, properties, etc.

For vectors \vec{n} of other terms, we will form a set of dictionaries of synonymous terms mentioned on a given site, $O' \cup O''$, on the basis of which we will form vectors \vec{O}', \vec{O}'' , such that

 $O_i = \begin{cases} 0, o \notin \vec{n} \\ 1, o \in \vec{n} \end{cases}$, whence, the criterion for the proximity of vectors takes the form (2):

$$F_{tertiary} = \sum_{j=0}^{J} (O'_j - O''_j)^2 \rightarrow \min .$$
⁽²⁾

Based on the coefficient $F_{tertiary}$, α can be obtained (3):

$$\alpha = \frac{\rho(\vec{n}_U)}{F_{tertiary}},\tag{3}$$

where a vector $\dot{\vec{n}}_U = \dot{\vec{n}}' \cup \dot{\vec{n}}''$, but $\rho(\ddot{\vec{n}}_U)$ – vector's length.

The meaning of the coefficient α is that: the greater the degree of similarity according to the tertiary criterion, the more likely that the terms describe the same product or service, the less stringency is required for similarity according to the primary and secondary criteria. However, the use of only these two criteria for assessing the degree of similarity of terms is not enough, since quite often, the information about the product is incomplete.

For greater accuracy in assessing the degree of closeness of terms, we introduce a vector \vec{n} and a closeness criterion based on it in order to solve the problem of incompleteness of product data in web content. The vector \vec{n} is formed based on \vec{n} of the collected information about the subject area of the industry; in particular, it is taken into account, which counterparties in which markets operate with which goods, etc. For the restored vectors, a criterion similar to the primary one can be calculated (4):

$$F_{2}(\ddot{\vec{n}}) = (1 + \sum_{j=0}^{J} (m'_{j} - m''_{j})^{2})(1 + \sum_{j=0}^{J} (l'_{j} - l''_{j})^{2})(1 + \sum_{j=0}^{J} (g'_{j} - g''_{j})^{2}) \to \min .$$
(4)

From where, combining all three criteria into one (1-4), you can get a metric criterion for the proximity of terms (5).

Criterion (5) is interpreted as follows:

- $F_1 \le \alpha$ means that the terms should be similar according to the primary criterion. The larger vector is, the less they are similar according to the tertiary criterion;
- $F_2 \le \alpha \cdot F_1$ means that the similarity for the reconstructed vector \vec{n} should be greater (due to the large number of coordinates) than for the original vector \vec{n} within the error α .

$$F = \begin{cases} p' = p'', \\ |d' - d''| \le d_{treshold}, \\ F_1 \le \alpha, \\ F_1 \le \alpha \cdot F_2 \end{cases}$$
(5)

The algorithm for identifying a cluster of terms describing a unique kernel in the web stream is shown in Figure 2 [7].

As a result, based on three vectors that describe the semantic core and the introduced proximity metrics, it is possible to obtain a single semantic core of the web resource as a whole. However, it should be understood that the core is formed on a specific date and time. In addition, there is a kernel retrospective.

Let us first consider how the algorithm for constructing the semantic core of a web resource will function, taking into account the fact that in the web content itself there may be several core options. Moreover, if the content management system contains several options for web content, therefore, the number of kernel options increases.

3. Proposed algorithm

For each term (figure 2), a set of potential duplicates is formed among the terms included in the same category (product, time and geography). After that, within the framework of one set of duplicates, the terms are checked in pairs for similarity. In case two terms are not similar, then the compared term is excluded from the list of potential duplicates. Otherwise, the vectors of the two terms are combined into a new vector based on the following rule: if two terms describe one product or service and differ slightly, then the vectors combine these terms to a more accurate description of a product or service.

After considering one term within its set of duplicates, in case a cluster is selected, all vectors of terms included in the cluster are replaced by a cluster vector, which eliminates the formation of duplicates.

Highlighting terms in a story will differ from searching for a unique term among duplicates, only by the rule of checking the time of appearance of a term in web content. Obviously, the dates should be consistent within the same margin of error.



Figure 2: Algorithm

At the same time, the question of highlighting story chains in the web content of a certain industry is open: even if due to one important event, a number of less important derivative events occurred. To form a semantic core, it is easier to analyze the entire set of terms, rather than highlight a weighty primary source.

The accuracy of the approach largely depends on the quality of the primary processing of web content: the quality of syntactic models [3] and dictionaries.

Thus, for an error-free combination of terms in a category, it is necessary to use the developed algorithm and criterion for assessing the degree of proximity.

4. Software designing

Based on the requirements (figure 2), we start to design a software taking into account that there are two main users: user and administrator, which share the functionality. Using UML [7], the following use case diagram was designed (figure 3).



The main idea is to represent the final result of the algorithm execution in the form of an RDF schema. It is a handy tool that includes the basic components of a vector $\dot{\vec{n}}$. The operating principle of the software is shown in the diagrams - Figures 4 and 5.

As we can see, the software takes the web content of the web resource as the initial information. Next, the primary processing of web texts is performed (lemmatization and normalization). Then our clustering algorithm is applied and, as a result, an xml file is formed - an RDF schema.

5. Results

Let us look at the potential results of applying the proposed algorithm on the example of web content on the topic of astrology and psychology. Figure 6 shows an example of such content, followed by highlighting a vocabulary of terms based on the frequency of their occurrence in the text.

The web content of this site has changed three times for the last ten years. As you can see, from Google Analytics data, each such change was accompanied by a surge in user activity. In the example (figure 7), the semantic kernel was analyzed at the stage of its first change. The main task of the kernel was to fix in the minds of potential users of the site and the Internet in general that the CelestialTiming site [8] logo is associated with the direction of astrology and psychology. Subsequent changes to the core were aimed at the next two stages of the marketing of this web project.

The first stage is to promote the services of this website, namely the construction of a psychological portrait based on the user's personal data.

The second stage is to promote a new service - the school of psychology and astrology. In accordance with the stages, the semantic core of the website is also changed.





Figure 5: Final web form

Unfortunately, the creators of the web project made several mistakes. The first is that the kernels must be linked when changed. The second mistake is that the kernel should be changed not only on the website itself, but also on friendly links pointing to this kernel. The third mistake is that each core has its own life cycle and it is different from each other. This can be seen, for example, in figure 7. The fourth mistake is that the kernels, based on their aging effect, need changing more often than the creators of this project do. The fifth mistake is that all kernel changes must fit within the framework of a single marketing strategy.

All these comments were passed on to the developers of this web project.







Figure 7: Google analytics data

6. Future work

Subsequent research on this topic includes:

• development of an alternative approach to the description of the semantic kernel. A promising idea is to represent the kernel in matrix form using the principles of permutations based on expert assessments of relationships between terms. In this case, the matrix contains the same elements as the search engine matrix. Then it is possible to use a vector-space model to determine the proximity of nuclei. In addition, it is promising to use a genetic algorithm to obtain an optimal semantic kernel, on a set of existing or promising ones;

• development and testing of software in the Javascript language on the WordPos platform [9]. It is required to design this software as a component with the possibility of its subsequent integration into existing content management systems such as Wordpress and Opencart;

• also a promising area of research is the analysis of profiles in social networks in order to identify semantic cores and, on their basis, search for potential buyers of a product or service.

7. References

- [1] S. Orekhov, M. Godlevsky, O. Orekhova, Theoretical fundamentals of search engine optimization based on machine learning in: Proceesings of the 13th International Conference on ICT in Education, Research and Industrial Applications. Integration, Harmonization and Knowledge Transfer, ICTERI '2017, CEUR-WS, 2017, Volume 1844, pp. 23-32
- [2] Y. Goldberg. Neural Network Methods in Natural Language Processing (Synthesis Lectures on Human Language Technologies). Morgan&Claypool Publ. USA, 2017
- [3] S. Orekhov, H. Malyhon, T. Goncharenko, Using Internet News Flows as Marketing Data Component in: Proceedings of the 4th International Conference on Computational Linguistics and Intelligent Systems. Volume 1: Main Conference, COLINS '2020, CEUR-WS, 2020, Volume 2604, pp. 358-373
- [4] M. Berry. Survey of Text Mining: Clustering, Classification and Retrieval. Springer, USA, 2004
- [5] E. Krupka, N. Tischby. Generalization from Observed to Unobserved Features by Clustering. Journal of Machine Learning Research, 2008, Volume 9, pp. 339-370
- [6] A. Jain, M. Murty, P. Flynn. Data clustering: A review. ACM Computing Surveys, 1999, Volume 31, No. 3, pp. 264–323
- [7] B. Rumpe. Agile modeling with UML. Springer, Germany, 2017
- [8] Web content source: Psychological self portrait is the path of self discovery based on celestial timing / Celestialtiming.com, 2021
- [9] J. Lengstorf, K. Wald. Pro PHP and jQuery. APress, USA, 2016