ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

INFORMATION TECHNOLOGY

UDC 004.9 **DOI: 10.20998/2079-0023.2022.01.09**

S. V. OREKHOV, H. V. MALYHON, N. K. STRATIIENKO

PROBLEM OF CLASSIFICATION OF SEMANTIC KERNELS OF WEB RESOURCE

The article presents a new theoretical basis for solving the problem of situational management of semantic cores identified on the basis of WEB content. Such a task arises within the framework of a new phenomenon called virtual promotion. Its essence lies in the fact that a real product can exist in two realities: online and offline. According to marketing theory, the lifetime in two realities is the same. However, in the online mode, the goods exist independently and in accordance with the laws of the use of Internet technologies. Therefore, based on the concept of a marketing channel, it was proposed to consider a message in such a channel as a semantic core. The core is a specially selected set of keywords that briefly describe the product and the corresponding need. It has been proposed that each need forms a so-called class of need. Therefore, the product description will either belong to this class or not. In addition, a product can be described by a different set of keywords, which means that different descriptions of the same product or several products, if there are any for sale in the enterprise, will fall into the demand class. As a result, in this work, it was proposed to consider the center of this class as the so-called *K*-candidate. It is the *K*-applicant that will be the semantic core that will be considered at the current iteration of the situational management process. In addition, in order to move from one situation to another, in other words, from one core to another, it is required to have such an alternative core. It can be safely taken either from the neighborhood of the need class center (*K*-applicant), or the center of another class (another *K*-applicant), if the product can cover several needs of a potential buyer. Then the actual task is to classify the classes of needs based on the text corpus in HTML format. Having a text corpus at the first stage, the task of synthesizing semantic cores is realized, and then the classification task itself. This article proposes the formulation of the classification problem, taking into account the features that the Internet technologies contribute to search engine optimization. In particular, it is proposed to use four metrics from the category of WEB statistics. And then it is proposed to use the clustering method to identify classes of needs, taking into account the fact that the *K*-applicant is presented as a semantic network or as a graph.

Keywords: semantic kernel, keyword, Ford – Fulkerson method, *K*-applicant.

С. В. ОРЄХОВ, Г. В. МАЛИГОН, Н. К. СТРАТІЄНКО

ЗАДАЧА КЛАСИФІКАЦІЇ СЕМАНТИЧНИХ ЯДЕР ВЕБ РЕСУРСУ

У статті представлено нову теоретичну базу для вирішення задачі ситуаційного управління семантичними ядрами, виділеними на основі ВЕБ контенту. Таке завдання виникає у рамках нового феномена під назвою віртуальне просування. Суть його полягає в тому, що реальний товар може існувати у двох реальностях: онлайн та офлайн. Відповідно до теорії маркетингу час життя у двох реальностях одне й теж. Однак у режимі онлайн товар існує самостійно і згідно із законами застосування Інтернет технологій. Тому в роботі на основі концепції маркетингового каналу було запропоновано розглядати повідомлення у такому каналі як семантичне ядро. Ядро є спеціально виділене безліч ключових слів, які коротко описують товар та відповідну йому потребу. Було запропоновано, кожна потреба формує так званий клас потреби. Отже, опис товару або належатиме даному класу чи ні. З іншого боку, товар можна описати іншим набором ключових слів, отже у клас потреби потраплять різні описи однієї й тієї ж товару чи кількох товарів, якщо такі є для підприємства продажу. В результаті в цій роботі було запропоновано вважати центр такого класу так званим *K*-претендентом. Саме *K*-претендент і буде тим семантичним ядром, яке на поточній ітерації процесу ситуаційного управління розглядатиметься. Крім того, для переходу від однієї ситуації до іншої, тобто від одного ядра до іншого, потрібно мати таке альтернативне ядро. Його можна сміливо брати або з околиці центроїду класу потреби (*K*-претендента), або центроїд іншого класу (інший *K*-претендент), якщо товар може покрити кілька потреб потенційного покупця. Тоді актуальне завдання класифікації класів потреб на основі текстового корпусу у форматі HTML. Маючи текстовий корпус першому етапі реалізується завдання синтезу семантичних ядер, та був власне завдання класифікації. У цій статті запропоновано постановку завдання класифікації з урахуванням особливостей, що вносять Інтернет технології, пов'язані з пошуковою оптимізацією. Зокрема, запропоновано використовувати чотири метрики з розряду ВЕБ статистики. І далі запропоновано використовувати метод кластеризації для виділення класів потреб з урахуванням того, що *K*-претендент представлений як семантична мережа або як граф.

Ключові слова: семантичне ядро, ключове слово, метод Форда – Фалкерсона, *K*-претендент.

Introduction. In the paper it is offered to consider a problem of classification of semantic kernels. In general, this problem is formulated as follows [1-2]. There is a set of objects X, and Y is a set of class numbers. Then we create a mapping: $F: X \to Y$, the value of which is known only on a given set of pairs $X_L = \{(x_1, y_1), ..., (x_m, y_m)\}$. This is a training sample. You need to build an algorithm $A: X \to Y$ that allows you to classify a new object $x \in X$. We adapt the classical formulation of this problem to our case, taking into account the already existing problem of situational management [3].

We will call a *K*-applicant such a kernel that launching it in the promotion channel leads to the sale of goods or services over the Internet. We will assume that the *K*-applicant forms the center of the cluster of kernels, i.e. it is a reflection or annotation of the main content of

the text describing the product or service for a potential buyer. Consider the main properties of the *K*-applicant.

K-applicant is a brief description of the product or the need that the product covers. It describes a product that should be purchased from the perspective of a potential buyer. Taking into account the theory and practice of search engine optimization on the Internet [4–7], *K*-applicant is a set of keywords that describes the set of keywords and the rules they form. The typical frequency of *K*-applicant in web content is 5–7 %. It must be reflected in the web content header.

In the paper it is considered that the *K*-applicant is a center of the class of semantic kernel. There can be as many such cluster center as you want, because each product can cover several needs. We will assume that one *K*-applicant presents one need that closes our product or service to the buyer. It is very important to describe the *K*-applicant in terms of quantitative indicators. We can generally talk about three main key indicators. The first indicator is the frequency of appearance in web text. The second is the number of keywords that make up the *K*-applicant. The third is the number of rules from the point of view of the theory of knowledge representation, which are hidden in the body of the *K*-applicant, if it can be represented as a semantic network.

Then *K*-applicant will be called an annotation of the description of the need that closes the product that we promote online. This annotation is presented in the form of a semantic network built by an algorithm [8]. Different *K*-applicants can be compared by the number of rules, the frequency of occurrence in web content, the frequency of occurrence in the search engine database and the number of keywords in its composition. Let's mark these indicators accordingly I_1, I_2, I_3, I_4. Then schematically the classification process can be depicted as follows – fig. 1.

Problem statement. The process shown in fig. 1 can be described as follows. Let the web content describe several needs that cover the product or service being promoted. Each of these needs is met by a set of keywords and rules that connect them to the appropriate content. This content is presented in the form of a semantic network and is called *K*-applicant. Variants of semantic networks, which include similar keywords with similar rules, will also be formed around the *K*-applicant.

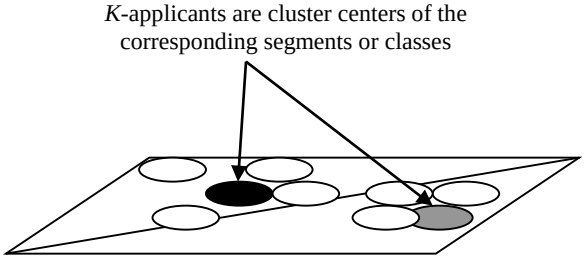

Fig. 1. Classification process based on *K*-applicants

That is, there are analogies around the *K*-applicant, other similar versions of semantic networks, which give the same meaning, but with different combinations of keywords and rules. Such a group of semantic networks (*K*-applicant and analogies) will be denoted as a class of needs. Then let there be a set of needs classes $Y_N = \{y_1, ..., y_n\}$ for each product or service being promoted. And the problem of classification is transformed into two problems: self-classification and clustering [9–10]. That is, you must first form a set of classes Y_N, and then assign to the class the existing semantic kernel.

Thus, the problem of classification of the semantic kernel is formulated as follows: having many classes of needs Y_N, we need to build a classifier $A: X \to Y$ that can assign the semantic kernel to the appropriate class of needs. In general, the classifier can be organized as a linear, DNF (disjunctive normal form) rule or neural network [11–12]. But given the fact of solving the problem of nuclear fusion, we can say that the problem of classification first degenerates into the problem of clustering of semantic kernel. This effect occurs because we synthesize a set of kernels, and then we have to divide this set into classes of need Y_N. If we re-synthesize kernel, the problem of classification is already solved.

The problem of kernel clustering is formulated as follows (fig. 1). We believe that we have many semantic kernels. There is also a function of the distance between two kernels. Then it is necessary to divide the set into subsets that do not intersect and which include nuclei that are close in metric.

Having many problems of the dissertation, we will consider approaches to their solution based on existing methods of artificial intelligence, machine learning, soft computing and data acquisition technology.

Proposed approach. In the paper to solve the clustering problem it is proposed to use the Ford – Fulkerson algorithm [9]. Consider formally its stages.

At the input of the algorithm we submit a semantic network $SN = \{V, E\}$, indicating the source *s*, the drain *t* and the bandwidth matrix *c*. Moreover, the elements of the matrix *c* will be calculated by formula (1). The result of the algorithm is the maximum possible flow *f* from *s* to *t*.

$$d(i, j) = \sqrt{\sum_{k=1}^{4} (I_k^i - I_k^j)^2} \quad (1)$$

.

If there is no connection between the vertices, the distance is considered infinite. In addition, we will introduce a capacity indicator:

$$c(i, j) = \frac{1}{d(i, j)} \quad (2)$$

.

The algorithm includes the following steps:

Step 1. We believe that $f(i, j) = 0$ for everyone $(i, j) \in E$. The final network coincides with the initial: $N = \overline{N}$.

Step 2. Choose any path $p \in \overline{N}$ from *s* to *t* such that $c(i, j) > 0$ for all edges $(i, j) \in p$, and go to step 2. If such a path does not exist, the algorithm ends.

Step 3. In the path found in the first stage p we find the edge with the minimum bandwidth $c_{\min}(i, j)$ and go to the third step.

Step 4. For each edge in the found way we increase a stream on size $c_{\min}$, that is we consider $f(i, j) = f(i, j) + c_{\min}$ for all $(i, j) \in p$; the flow for the edges of the opposite path is reduced by the value $c_{\min}$, i.e. for all $f(j,i) = f(j,i) - c_{\min}$. Let us move on to the fourth step.

Step 5. Adjust the final network $\overline{N}$. For all edges in the found path and for the opposite edges we calculate a new bandwidth. Remove the edge with zero bandwidth, and with non-zero – add to the final network and go to the first step.

For the numerical implementation of the algorithm, we introduce the definition of the stop function of the algorithm as [13]:

$$Q = \frac{r}{R}, \tag{3}$$

$$r = \frac{1}{K}\frac{1}{\left|P'_K\right|}\sum_{(i,j)\in P'_K} d(i, j),$$

$$R = \frac{1}{K}\frac{1}{\left|P'_K\right|}\sum_{(i,j)\in P'_K} d(i, z_K),$$

where z_K – class center K, P'_K – set of keywords from semantic network which describes the semantic kernels of web content.

Future work. The solution of the problem of classification of semantic kernels is a part of the task of situational management. As it was said, for its solution it is required to allocate classes of needs. Therefore, the directions for further research are alternative methods for separating classes and cluster centers of these classes. Classification methods based on soft computing and graph theory are promising.

Summary. In this article, such new scientific results have been reviewed to complete the task of classification, as well as: 1) the statement of the task of classifying semantic cores on the basis of seeing the classes of needs on the basis of semantic measures was formulated, as it was generated on the basis of web content; 2) got further development the method of solving the problem of classification; 3) previously proposed a change of metrics for visualizing consumer classes based on web content that describes a product for sale on the Internet.

References

1. Aggarwal C. C., Zhai C. X. *A survey of text classification algorithms*. *Mining Text Data*. Берлин: Springer Science-Business Media LLC, 2012. С. 163–222.
2. Остапец А. А. Решающие правила для ансамбля из цепей вероятностных классификаторов при решении задач классификации с пересекающимися классами. *Машинное обучение и анализ данных*. Москва: МФТИ, 2016. Том 2, №3. С. 276–285.
3. Поспелов Д. А. *Ситуационное управление: теория и практика*. Москва: Наука, 1986. 288 с.
4. Нееловой Н. М. *Энциклопедия поискового продвижения Ingate*. Москва: ИП Андросов, 2017. 541 с.
5. Бролина А. М. *Контекстная реклама: профессиональный апгрейд для увеличения продаж. Практикум от экспертов*. Москва: ООО «Ингейт Реклама», 2015. 44 с.
6. Sharma U., Thakur K. S. A Study on Digital Marketing and its Impact on Consumers Purchase. *International Journal of Advanced Science and Technology*. США: TUC, 2020. №29(3). С. 13096 – 13110.
7. García J., Lizcano D., Ramos C., Matos N. Digital Marketing Actions That Achieve a Better Attraction and Loyalty of Users: An Analytical Study. *Future Internet*. Щвейцария: MDPI, 2019. №11(130). С. 1–16.
8. Godlevsky M., Orekhov S., Orekhova E. Theoretical Fundamentals of Search Engine Optimization Based on Machine Learning. *CEUR WS*. США, 2017. № 1844. С. 23–32.
9. Коннов И. В., Кашина О. А., Гильманова Э. И. Решение задачи кластеризации методами оптимизации на графах. *Ученые записки казанского университета. Серия физико-математические науки*. Казань: КПФУ, 2019. Т. 161, кн. 3. С. 423–437.
10. Осипенко В. В. Два підходи до розв'язання задачі кластеризації у широкому сенсі з позицій індуктивного моделювання. *Енергетика і автоматика*. Київ: НУБПУ, 2014. №1. С. 83–97.
11. Khan A., Baharudin B., Lee L., Khairullah K. A Review of Machine Learning Algorithms for Text-Documents Classification. *Journal of advances in information technology*. 2010. Том 1, № 1. С. 4–20.
12. Неделько В. М. Исследование эффективности некоторых линейных методов классификации на модельных распределениях. *Машинное обучение и анализ данных*. Москва: МФТИ, 2016. Том 2, №3. С. 305–328.
13. Сивоголовко Е. Методы оценки качества четкой кластеризации. *Компьютерные инструменты в образовании*. Санкт-Петербург: ЛЭТИ, 2011. № 4. С. 14–31.

References (transliterated)

1. Aggarwal C. C., Zhai C. X. *A survey of text classification algorithms*. *Mining Text Data*. Berlin: Springer Science-Business Media LLC Publ., 2012, pp. 163–222.
2. Ostapez A. A. Reshayuschie pravila dlya ansamblya iz zepey veroyatnostnuh klassifikatorov pri reshenii zadach klasifikazii s peresekayuschimisya klassami. [Decision rules for an ensemble of chains of probabilistic classifiers in solving classification problems with intersecting classes] *Machine learning and data analysis*. Moscow: MFTI Publ., 2016, vol. 2, no. 3, pp. 276–285.
3. Pospelov D. A. *Situazionnoye upravlenie: teoriya i praktika*. [Situational management: theory and practice]. Moscow: Nauka Publ., 1986. 288 p.
4. Neelova N. M. *Enziklopaediya poiskovogo prodvizeniya Ingate*. [Encyclopedia of Search Engine Promotion Ingate]. Moscow: IP Androsov Publ., 2017. 541 p.
5. Brolina A. M. *Kontextnaya reklama: profesionalnuy upgrate dlya uvelicheniya prodaz. Praktikum ot expertov*. [Contextual advertising: a professional upgrade to increase sales. Workshop from experts]. Moscow: OOO «Ingate Reklama» Publ., 2015. 44 p.
6. Sharma U., Thakur K. S. A Study on Digital Marketing and its Impact on Consumers Purchase. *International Journal of Advanced Science and Technology*. 2020, no. 29(3), pp. 13096–13110.
7. García J., Lizcano D., Ramos C., Matos N. Digital Marketing Actions That Achieve a Better Attraction and Loyalty of Users: An Analytical Study. *Future Internet*. Switzerland: MDPI Publ., 2019, no. 11(130), pp. 1–16.
8. Godlevsky M., Orekhov S., Orekhova E. Theoretical Fundamentals of Search Engine Optimization Based on Machine Learning. *CEUR WS*, USA, 2017, vol. 1844, pp. 23–32.
9. Konnonov I. V., Kashina O. A., Gilmanova E. I. Reshenie zadachi klasterizazii metodami opimiozazii на графах. [Solving the clustering problem by optimization methods on graphs]. *Scientific notes of Kazan University. Series of physical and mathematical sciences*. Kazan: KPFU Publ., 2019, vol. 161, book 3, pp. 423–437.

10. Osipenko V. V. Dva pidhodu do rozvajannya zadachi klasterizazii u shirokomy sensi z pozuzii induktuvnogo modelyuvannya. [Two approaches to solving the problem of clustering in a broad sense from the standpoint of inductive modeling]. *Energy and automation*. Kyiv: NUBPU Publ., 2014, no. 1, pp. 83–97.
11. Khan A., Baharudin B., Lee L., Khairullah K. A Review of Machine Learning Algorithms for Text-Documents Classification. *Journal of advances in information technology*. USA, 2010, vol. 1, no. 1, pp. 4–20.
12. Nedelko V. M. Issledovaniye efektovnosti nekotoruh lineynuh metodov klassifikazii na modelnuh raspredeleniyah.[Investigation of the efficiency of some linear classification methods on model distributions]. *Machine learning and data analysis*. Moscow: MFTI Publ., 2016, vol. 2, no. 3, pp. 305–328.
13. Sivogolovko E. Metodu ozenki kachestva chetkoy klasterizazii. [Methods for assessing the quality of clear clustering]. *Computer tools in education*. SPb.: LETI Publ., 2011, no. 4, pp. 14–31.

Received 23.04.2022

Відомості про авторів / About the Authors

Орєхов Сергій Валерійович – кандидат технічних наук, доцент, Національний технічний університет «Харківський політехнічний інститут», доцент кафедри програмної інженерії та інтелектуальних технологій управління; м. Харків, Україна; ORCID: https://orcid.org/0000-0002-5040-5861; e-mail: sergey.v.orekhov@gmail.com

Малигон Геннадій Васильович – аспірант, Національний технічний університет «Харківський політехнічний інститут», аспірант кафедри програмної інженерії та інтелектуальних технологій управління; м. Харків, Україна; ORCID: https://orcid.org/0000-0001-5448-2488; e-mail: gmalygon@gmail.com

Стратієнко Наталія Костянтинівна – кандидат технічних наук, доцент, Національний технічний університет «Харківський політехнічний інститут», професор кафедри програмної інженерії та інтелектуальних технологій управління; м. Харків, Україна; ORCID: https://orcid.org/0000-0002-7925-6687; тел. (057)707-64-74, email: strana.snk@gmail.com

Orekhov Sergey Valerievich – Candidate of Technical Sciences (PhD), Docent, National Technical University «Kharkov Polytechnic Institute», Accosiate Professor of Software Engineering and Management Intelligent Technologies department; Kharkov, Ukraine; ORCID: https://orcid.org/0000-0002-5040-5861; e-mail: sergey.v.orekhov@gmail.com

Malyhon Hennadiy Vasilievich – Postgraduate Student, National Technical University «Kharkov Polytechnic Institute», Postgraduate Student of Software Engineering and Management Intelligent Technologies department; Kharkov, Ukraine; ORCID: https://orcid.org/0000-0001-5448-2488; e-mail: gmalygon@gmail.com

Stratiienko Nataliia Kostiantunivna – Candidate of Technical Sciences (PhD), Docent, National Technical University «Kharkiv Polytechnic Institute», Professor at the Software Engineering and Management Intelligent Technologies Department; Kharkiv, Ukraine; ORCID: https://orcid.org/0000-0002-7925-6687; tel. (057)707-64-74, email: strana.snk@gmail.com