МАТЕМАТИЧНЕ І КОМП'ЮТЕРНЕ МОДЕЛЮВАННЯ

МАТЕМАТИЧЕСКОЕ И КОМПЬЮТЕРНОЕ МОДЕЛИРОВАНИЕ

MATHEMATICAL AND COMPUTER MODELING

UDC 004.9 **DOI: 10.20998/2079-0023.2021.01.08**

S. V. OREKHOV, H. V. MALYHON

MODELLING SEMANTIC KERNEL OF WEB RESOURCE

The article presents an attempt to describe mathematically the effect of the semantic kernel of a web resource on the Internet. In accordance with the theory of marketing, the product that we want to sell on the network is characterized by the following basic properties: price, time and place. In other words, a potential buyer wants to receive a given product in the right place at a given time. To satisfy this need, it is necessary to use the classic component of marketing, product promotion. However, this component is now becoming a fully virtual instrument. This tool functions in a hypertext, video and image environment. Therefore, the user analyzes the meaning of these elements in order to get the desired product. The results of web projects carried out in this area indicate the emergence of a new phenomenon, which reflects the main meaning of virtual promotion – this is the semantic core. The core is a short annotation of the main properties of the product, its location and time of appearance. Therefore, the purpose of this article is both a presentation of a new object of research and a mathematical description. It is assumed that the semantic core is formed on the basis of natural language terms. In other words, the semantic core is a set of keywords that are grouped by meaning. We propose to use data mining approaches for clustering to group terms. The classic clustering method at the moment is *k*-means. The article presents a model of the semantic core based on this method. This method and its distance function are considered as the second stage of web content processing. At the first stage, web content is converted into a semantic web. However, the *k*-means technique has significant drawbacks when modeling the semantic core. Therefore, in the development of this idea, the work shows an alternative way to modeling the kernel. As an alternative approach, the construction of clusters based on the concept of maximum flow is considered. This approach has the significant advantage that the type of links in the semantic network overlaps with the type of distance function in this method. As a result, on a real web project, the effect of the connection between the semantic core model and the level of new users of the web resource was demonstrated over the past five years.

Keywords: semantic kernel, keyword, *k*-means, max flow.

С. В. ОРЄХОВ, Г. В. МАЛИГОН

МОДЕЛЮВАННЯ СЕМАНТИЧНОГО ЯДРА ВЕБ РЕСУРСУ

У статті представлена спроба описати математично ефект семантичного ядра веб ресурсу в середовищі Інтернет. Відповідно до теорії маркетингу продукт, який ми бажаємо продати в мережі, характеризуються такими основними властивостями: ціна, час і місце. Іншими словами потенційний покупець бажає отримати заданий товар в потрібному місці в заданий час. Щоб задовольнити цю потребу, треба використовувати класичну компоненту маркетингу просування товару. Однак зараз ця компонента стає повністю віртуальним інструментом. Цей інструмент функціонує в середовищі гіпертекстів, відео та зображень. Тому користувач аналізує зміст даних елементів, щоб отримати бажаний товар. Результати виконаних в цій області веб проектів свідчать про появу нового явища, яке відображає основний зміст віртуального просування – це семантичне ядро. Ядро являє собою коротку анотацію основних властивостей товару, його місце розташування і час появи. Тому метою даної статті є як презентація нового об'єкта дослідження, так і математичний опис. Передбачається, що семантичне ядро формується на основі термінів природної мови. Іншими словами семантичне ядро – це безліч ключових слів, які згруповані за змістом. Ми пропонуємо використовувати для угрупування термінів підходи технології Data mining по кластеризації. Класичним методом кластеризації на даний момент є *k*-середніх. У статті представлена модель семантичного ядра на основі даного методу. Цей метод і його функції дистанції розглядаються як другий етап обробки веб контенту. На першому етапі веб контент конвертується в семантичну мережу. Однак методика *k*-середніх має суттєві недоліки при моделюванні семантичного ядра. Тому в розвитку даної ідеї в роботі показаний альтернативний шлях до моделювання ядра. В якості альтернативного підходу розглядається побудови кластерів на основі концепції максимального потоку. Цей підхід має істотну перевагу, яке полягає в тому, що тип зв'язків в семантичної мережі перегукується з типом функції дистанції в даному методі. В результаті на реальному веб проект продемонстрований ефект зв'язку між моделлю семантичного ядра і рівнем нових користувачів веб ресурсу протягом останніх п'яти років.

Ключові слова: семантичне ядро, ключове слово, метод *k*-середніх, максимальний потік.

С. В. ОРЕХОВ, Г. В. МАЛЫГОН

МОДЕЛИРОВАНИЕ СЕМАНТИЧЕСКОГО ЯДРА ВЕБ РЕСУРСА

В статье представлена попытка описать математически эффект семантического ядра веб ресурса в среде Интернет. В соответствии с теорией маркетинга продукт, который мы желаем продать в сети, характеризуются следующими основными свойствами: цена, время и место. Другими словами, потенциальный покупатель желает получить заданный товар в нужном месте в заданное время. Чтобы удовлетворить данную потребность, надо использовать классическую компоненту маркетинга продвижение товара. Однако сейчас эта компонента становится полностью виртуальным инструментом. Этот инструмент функционирует в среде гипертекстов, видео и изображений. Поэтому пользователь анализирует смысл данных элементов, чтобы получить желаемый товар. Результаты выполненных в этой области веб проектов свидетельствуют о появлении нового явления, которое отражает основной смысл виртуального продвижения – это семантическое ядро. Ядро

© S. V. Orekhov, H. V. Malyhon, 2021

представляет собой краткую аннотацию основных свойств товара, его местоположение и время появления. Поэтому целью данной статьи является как презентация нового объекта исследования, так и математическое описание. Предполагается, что семантическое ядро формируется на основе терминов естественного языка. Другими словами, семантическое ядро – это множество ключевых слов, которые сгруппированы по смыслу. Мы предлагаем использовать для группировки терминов подходы технологии Data mining по кластеризации. Классическим методом кластеризации на данный момент является *k*-средних. В статье представлена модель семантического ядра на основе данного метода. Этот метод и его функции дистанции рассматриваются как второй этап обработки веб контента. На первом этапе веб контент конвертируется в семантическую сеть. Однако методика *k*-средних имеет существенные недостатки при моделировании семантического ядра. Поэтому в развитии данной идеи в работе показан альтернативный путь к моделированию ядра. В качестве альтернативного подхода рассматривается построения кластеров на основе концепции максимального потока. Этот подход имеет существенное преимущество, которое заключается в том, что тип связей в семантической сети перекликается с типом функции дистанции в данном методе. В результате на реальном веб проекте продемонстрирован эффект связи между моделью семантического ядра и уровнем новых пользователей веб ресурса на протяжении последних пяти лет.

Ключевые слова: семантическое ядро, ключевое слово, метод *k*-средних, максимальный поток.

Introduction. The semantic kernel is a new concept based on the assumption that the modern Internet environment is a semantic global network [1–2]. We assume that every web resource on the web is a particle or book in a global virtual library. The challenge is to find knowledge in such a library. But there is another problem, in what place and how to place our knowledge, for example, about a product or service, so that this knowledge is found, read and accepted by potential customers.

Our scientific work is based on the assumption that since the Internet is a semantic network, and in terms of artificial intelligence, it is a kind of knowledge base. Together with the search server, the Internet can then be regarded as a kind of expert system. Therefore, for this system to function, it must be trained.

According to the classics, the learning process takes place in a mode with a teacher and without him [3]. The second option is the most acceptable in this situation, since we kind of directly invest knowledge into the system, which actually happens when a programmer creates and places a web resource on the network.

To train an expert system, three elements are required: a training method, a stopping criterion, and a training sample. We propose to consider the semantic core of the web resource as a training sample for our global semantic network.

We will assume that the semantic kernel of a web resource is a short annotation in the form of a set of keywords. These words define the main meaning of the web content of the web resource in question.

Then there is an urgent problem of automatic formation of such a semantic kernel. To solve it, it is required to propose a method for modeling the semantic core in mathematical or algorithmic form.

Problem statement. Let's assume that web content contains several semantic cores. Why? The fact is that, as a rule, a web resource is created with the aim of promoting one or more goods or services. In addition, if the current core does not provide a sufficient level of sales, then they try to either replace it or change it. Therefore, the content management system of a web resource contains several versions of web content. This is a typical situation for many content management systems: Wordpress, Opencart, Drupal, ModX, and others.

Thus, we believe that the semantic core is a certain cluster that combines keywords according to some marketing sales strategy. Then the problem of constructing a semantic core is formulated as the problem of identifying an unknown number of clusters or a clustering problem when the finite number of clusters is unknown. In addition, we do not know the shape of the clusters, that is, semantic cores can contain a different number of keywords, and it is also possible to overlap.

Let us consider the k-means method as the basic algorithm for the formation of the semantic core. This approach meets almost all of the above requirements at the moment. In addition, it is convenient to consider it as the first attempt to solve the clustering problem. It is also convenient for software implementation.

Let it be $X = \{x_i\}$ a set of keywords from web content. $C = \{c_i\}$ is a set of cluster centers (semantic kernels). Then the measure of error:

$$E(X, C) = \sum_{i=1}^{n} \|x_i - c_i\|. \quad (1)$$

where c_i – closest to x_i cluster center.

In this case, the algorithm itself will include the following actions. The first step is to initialize the cluster centers $C = \{c_i\}$. We will assume that the centers of the clusters are those keywords (terms) that reflect either the names of goods and services, or related goods.

Until the belonging of the terms stops changing, we perform one of two actions. Determine the belonging of the term to the cluster by the formula (2). Next, determine the new center of the cluster by formula (3).

$$cluster_i = \arg\min_{c \in C} f(x_i, c). \quad (2)$$

where $f(x_i, c)$ – distance function to cluster center.

$$c = \frac{\sum_{j=1\ cluster_i=c} \hat{f}_j(x_i)}{\sum_{j=1\ cluster_i=c} 1}. \quad (3)$$

where $\hat{f}$ – the function of term definition as cluster center.

The main disadvantage of such an algorithm for constructing a semantic core is that the number of clusters must be known. However, this condition is not always feasible.

To fulfill this condition, we can use at the first stage the algorithm for constructing a semantic kernel proposed in [1]. As centers of clusters, we will take only those terms that have the maximum frequency of occurrence in the text, or that have the maximum number of links with other terms. There are two ways to define a function $\hat{f}$.

The function of the distance between the terms is also proposed to be selected on the basis of the semantic

network representing the web content. In this case, both the weight of the link and the number of edges between terms in the network can be taken into account. This will be the way to define the function $f(x_i, c)$.

Proposed approach. An alternative approach to constructing clusters of terms and, accordingly, semantic kernels will be the clustering technique based on the search for the maximum flow [4]. This approach does not require knowing the number of clusters in advance. In addition, the semantic network obtained at the first stage is a graph indicating, in essence, the weight of an edge. Therefore, it is proposed to use the following assumptions.

Let the semantic network $SN = \{V_s, E_s\}$ obtained using the algorithm [1] be given. V_s – the set of vertices of the network, and E_s accordingly the set of its edges. We will understand the distance $d(i,j)$ between two vertices i and j, as the weight of the edge that connects them. If there is no edge between the vertices, then the distance is infinite. Then the bandwidth of the edge (i,j) is represented as:

$$c(i,j) = \frac{1}{d(i,j)}. \quad (4)$$

We will consider a group of vertices with a beginning at the top s and end at the top t as the semantic kernel. This group of vertices is characterized by the maximum flow v. Stream this function, defined as follows:

$$f: E \to R^+. \quad (5)$$

The function (5) is being satisfied the conditions:

$$f(i,j) \le c(i,j). \quad (6)$$

$$\sum_{A(i)} f(i,j) - \sum_{B(i)} f(j,i) = \begin{cases} v, i = s \\ 0, i \notin \{s,t\}. \\ -v, i = t \end{cases} \quad (7)$$

$$A(i) = \{j \in V: (i,j) \in E\}. \quad (8)$$

$$B(i) = \{j \in V: (j,i) \in E\}. \quad (9)$$

To obtain the maximum flow, it is required to maximize the value under constraints (6)–(7). Then we will assume that the semantic kernel is a sub-network of the semantic network obtained by the algorithm [1], with the maximum flow in the sense of (5). In our case, the maximum flow should be interpreted as the maximum number of connections of the "isa" type. In other words, the maximum number of rules that bind terms in our semantic kernel [4].

The MaxFlow algorithm that implements this approach was presented in [4]. The paper proposes to apply this algorithm to the construction of a semantic kernel by clustering terms.

Future work. A significant drawback of the proposed approach to the construction of a semantic core is the insufficient consideration of the semantics of words and possible plot lines [5]. The advantage of the proposed approach is to take into account the relationship between the terms by applying the algorithm [1]. The resulting semantic network is further processed.

As a continuation of the research, one should focus on the application of the genetic algorithm [6–8] and the presentation of the process of constructing a kernel based on the methods of hierarchical clustering [9–12].

The fact is that the semantic network obtained at the initial stage contains edges with different types of links: "isa", "kindof" and "part of". These links can be interpreted as rules between terms corresponding to different types of distance function: single link, full link and medium link [9–12].

Summary. The above approach was applied to the analysis of the semantic kernel of one of the sites of the project for the presentation of a psychological portrait of a personality. To do this, the data from the Google Analytics service was compared with the constructed semantic kernels. It turned out that in this project, the semantic kernel was created once, and then changed in 2017–2018 (figure 1).

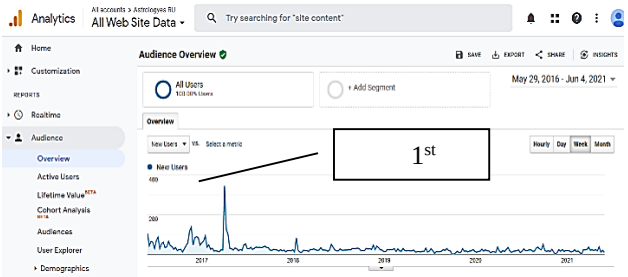

Fig. 1. Semantic kernel changes

As can be seen from the data in Figure 1, changing the kernel leads to a surge of interest in the web resource from new users. That is, the creation of a new version of the semantic kernel leads to the coverage of a certain group of users. Thus, the kernel reflects the interest of a given user group in this web content. Or in other words, this semantic kernel reflects the knowledge, wishes or preferences of a certain group of Internet users.

Consequently, this example confirms the assumption that the semantic kernel is a so-called training sample according to the assumption that in order to promote a product or service, we need to train the virtual space to recognize our web resource on the Internet.

Our research also shows how much the traffic of a web resource depends on such an effect as the aging of the semantic kernel [2].

References

1. Godlevsky M., Orekhov S., Orekhova E. Theoretical Fundamentals of Search Engine Optimization Based on Machine Learning. *CEUR–WS*. США, 2017. № 1844, С. 23–32.
2. Orekhov S., Malyhon H., Stratienko N., Goncharenko T. Software Development for Semantic Kernel Forming. *CEUR–WS*. США, 2021. № 2870. С. 1312–1322.
3. Amit K. *Artificial Intelligence and Soft Computing. Behavioral and Cognitive Modeling of the Human Brain*. США: CRC Press LLC, 2000. 788 с.
4. Коннов И. В., Кашина О. А., Гильманова Э. И. *Решение задачи кластеризации методами оптимизации на графах*. Ученые записи Казанского университета. Сер. Физ.-матем. науки. Казань: Казанский университет. 2019. Т. 161, кн. 3. С. 423–437.

5. Cherenkov I, Orekhov S. Approach for extracting events from news stream. *Eastern-European Journal of Enterprise Technologies*. 2013. Т. 1, № 4 (61), С. 62–64.
6. Preisach C., Burkhardt H., Schmidt-Thieme L, Decker R. *Data Analysis, Machine Learning and Applications*. Германия: Springer-Verlag Berlin Heidelberg, 2008. 703 с.
7. Han J., Kamber M, Pei J. *Data Mining Concepts and Techniques*. США: Morgan Kaufmann, 2012. 740 с.
8. Шумейко А. А., Сотник С. Л. *Интеллектуальный анализ данных. Введение в Data Mining*. Днепр: Белая Е. А., 2015, 223 с.
9. Уиллиамс У. Т., Ланс Д. Н. *Методы иерархической классификации Статистические методы для ЭВМ*. Москва: Наука, 1986. С. 269–301.
10. Venugopal K. R., Srinivasa K. G. и Patnaik L. M. *Soft Computing for Data Mining Applications*. Германия: Springer, 2009. 354 с.
11. Witten Ian H. , Frank E. *Data Mining. Practical Machine Learning Tools and Techniques*. США: Morgan Kaufmann, 2009. 558 с.
12. Дюран Б. *Кластерный анализ*. Москва: Статистика, 1977. 128 с.
13. Замятин А. В. *Интеллектуальный анализ данных: учебное пособие*. Томск: Издательский Дом Томского государственного университета, 2016. 120 с.

References (transliterated)

1. 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.
2. Orekhov S., Malyhon H., Stratienko N., Goncharenko T. Software Development for Semantic Kernel Forming. *CEUR WS*. USA, 2021. vol. 2870. pp. 1312–1322.
3. Amit K. *Artificial Intelligence and Soft Computing. Behavioral and Cognitive Modeling of the Human Brain*. USA: CRC Press LLC Publ., 2000. 788 p.
4. Konnov I. V., Kashina O. A., Gilmanova E. I. Reshenie zadachi klasterizazii metodamai optmizazii na grafah [*Solution of clusterization problem by graph optimization methods]*. Uchenye Zapiski Kazanskogo Universiteta. Seriya Fiziko-Matematicheskie Nauki [Scientific bulletin of Kazan University. Physical and Mathematical Series]. Kazan, Kazan University Publ., 2019, issue 161, no. 3, pp. 423–437.
5. Cherenkov I, Orekhov S. Approach for extracting events from news stream. *Eastern–European Journal of Enterprise Technologies*. 2013. vol. 1, no. 4 (61), pp. 62–64.
6. Preisach C., Burkhardt H., Schmidt-Thieme L, Decker R. *Data Analysis, Machine Learning and Applications*. Germany: Springer-Verlag Publ., 2008. 703 p.
7. Han J., Kamber M, Pei J. *Data Mining Concepts and Techniques*. USA: Morgan Kaufmann Publ., 2012. 740 p.
8. Shumeyko A.A., Sotnik S.L. *Intellektualnuy analyz dannuh. Vvedeniye v Data Mining* [Intelligent data analysis. Introduction in Data Mining]. Dnepr: Belaya Publ., 2015. 223 p.
9. Willams U. T., Lans D. N. Metodu ierarhicheskoy klassifikazii [Methods of hierarchical classification]. *Statistical methods of computer machine*. Moscow: Nauka Publ., 1986. pp. 269–301.
10. Venugopal K.R., Srinivasa K.G. and Patnaik L.M. *Soft Computing for Data Mining Applications*. Germany: Springer Publ., 2009. 354 p.
11. Witten Ian H., Frank E. *Data Mining. Practical Machine Learning Tools and Techniques*. USA: Morgan Kaufmann Publ., 2009. 558 p.
12. Duran B. *Klasternuy analiz* [Cluster analysis]. Moscow: Statistica Publ., 1977. 128 p.
13. Zamyatin A. V. *Intelektualnuy analiz dannuh. Uchebnoe posobie* [Intelligent data analysis: tutorial]. Tomsk: Tomsk state university Publ., 2016. 120 p.

Received 11.05.2021

Відомості про авторів / Сведения об авторах / About the Authors

Орєхов Сергій Валерійович – кандидат технічних наук, доцент кафедри програмної інженерії та інформаційних технологій управління; м. Харків, Україна; ORCID: https://orcid.org/0000-0002-5040-5861; e-mail: sergey.v.orekhov@gmail.com

Малигон Геннадій Васильович – аспірант кафедри програмної інженерії та інформаційних технологій управління; м. Харків, Україна; ORCID: https://orcid.org/0000-0001-5448-2488; e-mail: gmalygon@gmail.com

Орехов Сергей Валерьевич – кандидат технических наук, доцент, Национальный технический университет «Харьковский политехнический институт», доцент кафедры программной инженерии та информационных технологий управления; Харьков, Украина; ORCID: https://orcid.org/0000-0002-5040-5861; e-mail: sergey.v.orekhov@gmail.com

Малыгон Геннадий Васильевич – аспирант, Национальный технический университет «Харьковский политехнический институт», аспирант кафедры программной инженерии та информационных технологий управления; Харьков, Украина; ORCID: https://orcid.org/0000-0001-5448-2488; e-mail: gmalygon@gmail.com

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

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