Software Development for Semantic Kernel Forming

Sergey Orekhov, Henadii Malyhon, Nataliya Stratienko and Tetiana Goncharenko

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

Abstract

The article presents the results of the study of the semantic kernel forming process for a web resource. This study broadens the understanding of a new concept of the Semantic Web. It is based on four components, namely URI, ontology, data and semantic language. This concept, using a new tool – the semantic kernel, is implemented in the paper. Such a kernel is formed according to the principle of annotation, that is, it is a concentrated expression of the main meaning of a given web content. It is formed based on a unique algorithm relied on the semantic network and the method of Data mining technology. Thus, the work offers an alternative implementation of the components of the Semantic Web [1]. It is proposed to use the RDF schema as the semantic kernel representation format. Moreover, the software implementation of the entire approach is done in JavaScript using the Node JS library. The software was tested on the content of real web sites. The new effect of semantic kernel aging was detected for USA real web site.

Keywords 1

Semantic kernel, Semantic web, JavaScript, RDF, Data mining

1. Introduction

Before the buyer decides to buy the goods he needs, he usually compares alternatives looking for the information about goods (services) and analyzing them on the Internet. Based on the information received, the buyer makes a decision, while such factors as other people's attitudes, the benefits of the product, and other unforeseen factors influence the decision.

Therefore, the information left by users carries a marketing value. It is based on it and you can identify the events that in one way or another have a certain impact on the market situation. Nevertheless, this information appears as hypertexts, and to be more precise, a set of keywords. Therefore, it is an urgent task to investigate multiple keywords to evaluate their marketing value. We name the multiple keywords as a semantic kernel [2].

What is the semantic kernel actually? It is primarily a set of keywords that describe a briefly defined subject area. We will assume that such a semantic kernel describes a given product or service, that is, a textual display of a product or service in the virtual space. This display can be transmitted over the Internet and displayed on a web page. You can write a review for it, that is, tell your personal opinion about the product. In this case, the semantic kernel really acquires marketing knowledge (Figure 1).

Word or sentence	Demand	Proposal	Competitors	Partners	Ideas of content	Notes

Figure 1: Example of semantic kernel of a product

COLINS-2021: 5th International Conference on Computational Linguistics and Intelligent Systems, April 22–23, 2021, Kharkiv, Ukraine

EMAIL: sergey.v.orekhov@gmail.com (S. Orekhov); gmalygon@gmail.com (H. Malyhon); strana@kpi.kharkov.ua (N. Stratienko) ; tatianagoncharenko1806@gmail.com (T. Goncharenko)
ORCID: 0000-0002-5040-5861 (S. Orekhov); 0000-0001-5448-2488 (H. Malyhon); 0000-0002-7925-6687 (N. Stratienko); 0000-0001-6630-307X (T. Goncharenko)

However, using the kernel directly is inconvenient for the end user. In this scenario, the following problems arise. Firstly, it is not convenient for the user to search for the keywords that make up the kernel. Secondly, the keywords make sense to the user when there are connections between them. Thirdly, the kernel is changing over the time. And fourth, you need an understandable format for representing the kernel, both for the user and for the computer system.

Thus, it is required to propose a new form of representation of the semantic core that is convenient for understanding by a user, as well as with an acceptable data format that can be processed programmatically.

Therefore, the work proposes an integrated approach to the development of software for the formation of the semantic core and its research.

2. Problem statement

Recent studies have confirmed the fact that information flows on the Internet reflect market events [3]. However, these flows exist in the form of hypertexts. Moreover, if we perform a hypertext analysis of the content, then we have a set of keywords. It is the set that reflects market events; therefore, there is an urgent task to investigate the information flow in order to identify onboard keywords that reflect market events.

This task is complex and involves at least two tasks: the first task of highlighting a set of keywords (semantic kernel) and the second task of researching keywords to establish information about market events: participants, their actions, product prices etc.

A plurality of hypertext messages is the input. These messages have the following properties:

1. They consist of keywords.
2. All messages have an important feature – the moment of time.
3. Messages are formed on the basis of a glossary of terms (semantic core) about market events and their participants.
4. A user creates a series of messages using this dictionary.

Then, by analyzing the text, namely analyzing the content of the web page and using this dictionary, you can highlight a set of keywords depending on the time and probably the location.

We will call a set of keywords (semantic kernel) about the market goods a "card" of a product. This is a description of the product left by the user on the Internet. An example of such a product card for the auto parts and accessories market is presented in Figure 2 (http://celestialtiming.com).

We will refer to a set of keywords that describes information about the manufacturers of goods (market participants) as a card of a manufacturer.

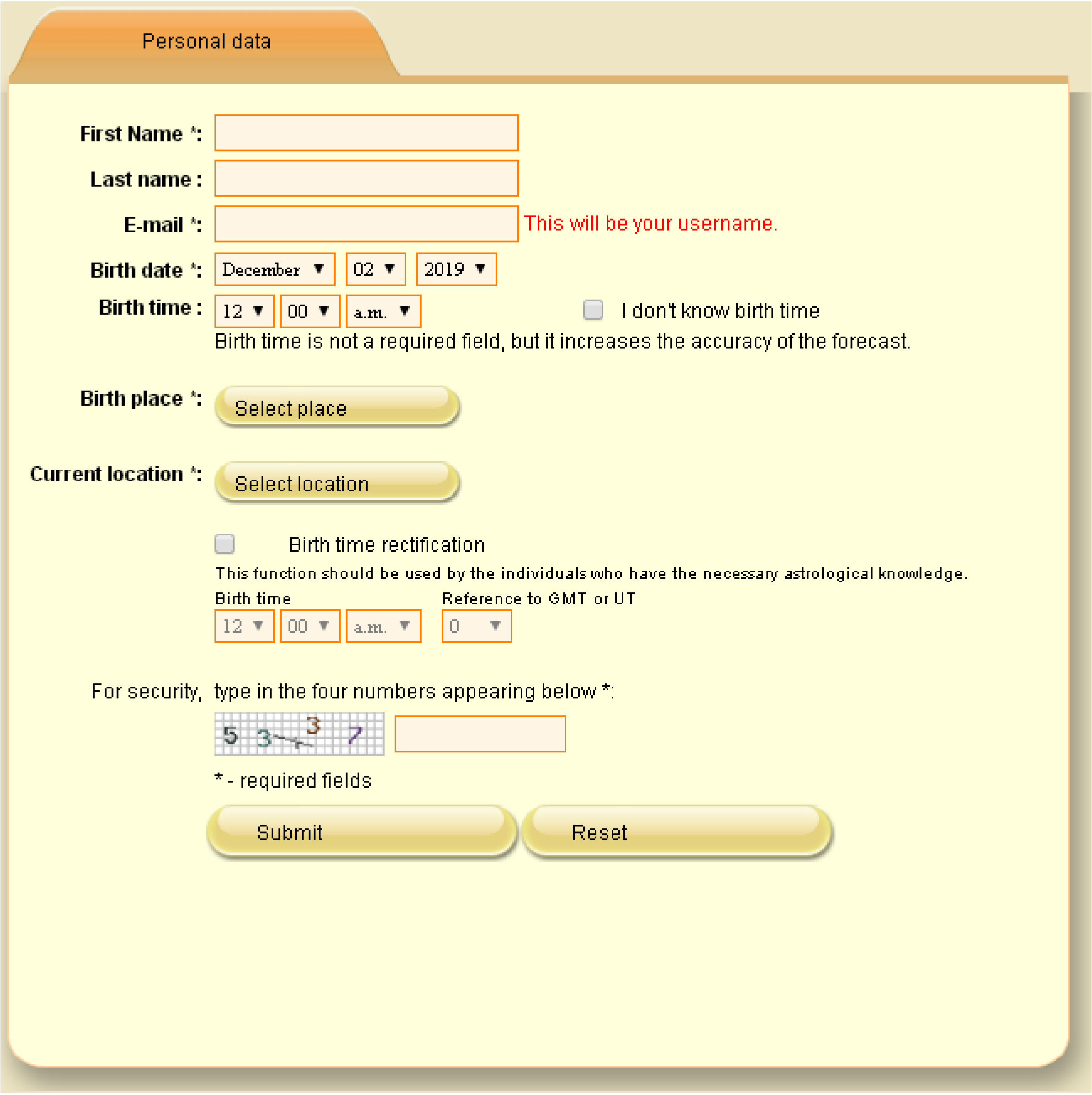
Then the tool for the accumulation and display of cards will be called a bulletin board of market participants about their products or services.

The bulletin board is an HTML page, where all the visitors of the site leave their ads, and all the visitors of the site can read them. The board then acts as a bridge between the product card, the manufacturer's card, and the potential or actual consumer of the product or service. In other words, the bulletin board is a repository of associative rules that are linked within the 4P principle: location (place), product, price, and advertising (promotion) [4-7]. Thus, the bulletin board becomes the main source of market events.

In addition, all the web pages of the board are subject to content analysis. Besides, to install a dictionary, you can use Data Mining technology [8] to identify multiple keywords.

This dictionary is dynamic and is mapped over the following periods: day, week or month. Therefore, using such a dictionary, you can determine the events with the help of a search service that led to a particular fact at a certain time.

Then the problem statement is formulated as follows: it is necessary to develop a software solution for the formation of a set of keywords (semantic kernel) on the example of the market hypertexts. To do a research it is offered to consider a set of key-words (semantic kernel). The set is dynamic, that is it is essential to build it depending on time. To form a semantic kernel, it is necessary to have: a dictionary, a product and a manufacturer cards, as well as a mechanism for establishing their relationship - a Bulletin Board.



Personal data

First Name *:

Last name :

E-mail *: This will be your username.

Birth date *: December ▼ 02 ▼ 2019 ▼

Birth time : 12 ▼ 00 ▼ a.m. ▼ □ I don't know birth time

Birth time is not a required field, but it increases the accuracy of the forecast.

Birth place *: Select place

Current location *: Select location

□ Birth time rectification

This function should be used by the individuals who have the necessary astrological knowledge.

Birth time: 12 ▼ 00 ▼ a.m. ▼

Reference to GMT or UT: 0 ▼

For security, type in the four numbers appearing below *:

* - required fields

Submit Reset

Figure 2: Real card of online service

In this study, we pay attention to the development of the software to form a semantic kernel based on the texts from Web page about a product card.

The task is defined as follows: it is necessary to develop software component to generate a semantic kernel based on a keyword dictionary. The semantic kernel is displayed in a form of a table or a xml document.

Schematically the business process in which such software will be involved is described using IDEF0 notation (see in Figure 3) [9].

From a user's point of view, the software performs the functions presented in the diagram of use cases – Figure 4 [9].

The following types of actors are distinguished for the software being developed: a user and a software developer or an administrator.

A user should have the following capabilities:

1. He enters a hypertext or a simple text to a special form.
2. He sees the semantic kernel.
3. He selects the view type of the semantic kernel representation.

A developer can integrate the component to existing software via setting functionality.

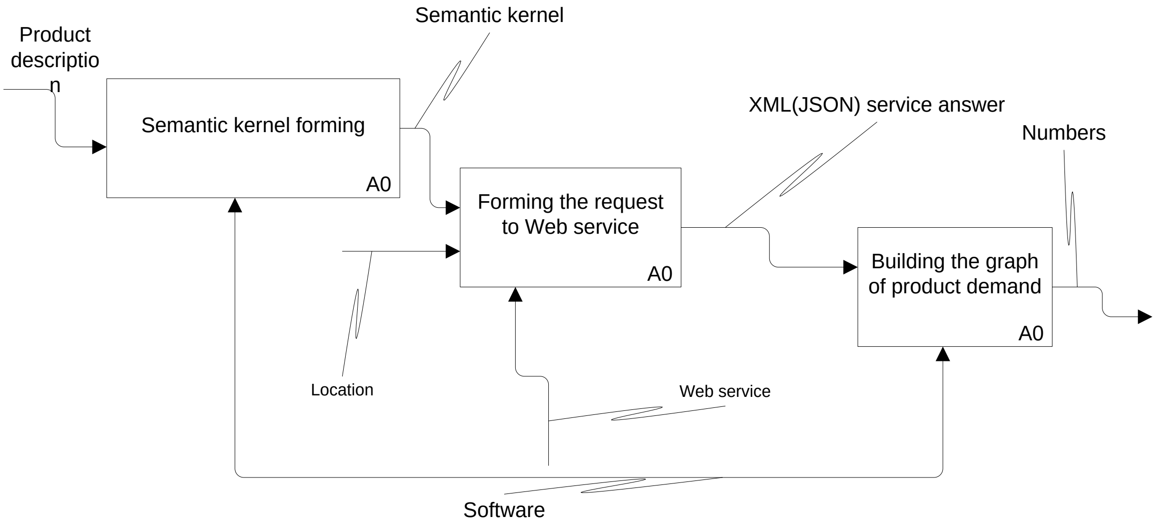


Figure 3: Business process decomposition

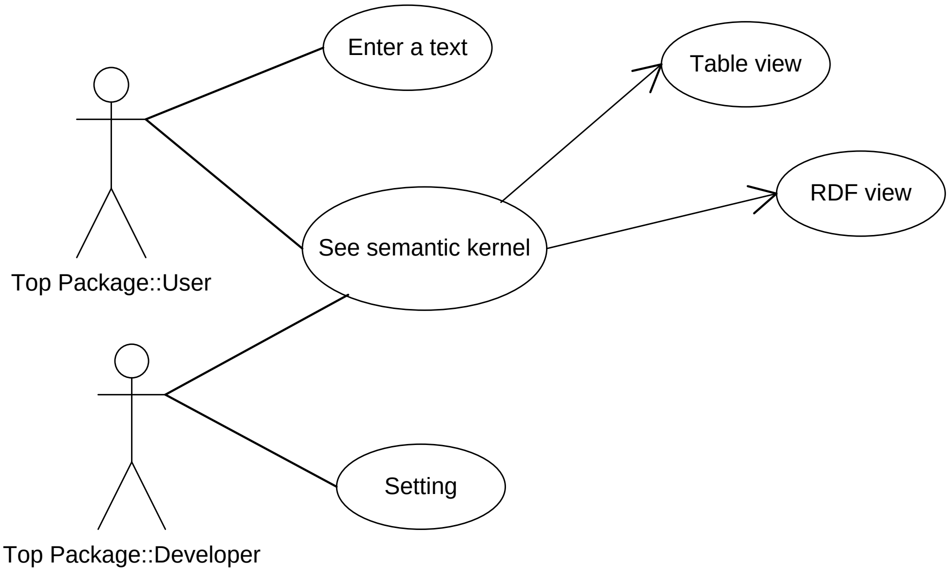


Figure 4: Business process decomposition

Non-functional requirements for the developed system as convenience of use and safety were highlighted. Convenience of use is understood as compliance of the user interface with standards, the convenience of work with it and its clarity. Security means protection against unauthorized access, the presence of registration and authorization.

To perform the task in the work, it is proposed:

- to consider existing ways of forming a semantic kernel;
- to propose solution schema;
- to design and test IT solution as a software.

3. Proposed methodology

In the current literature on search engines and services, the term "keyword dictionary or semantic kernel" is interpreted as the term of information search thesaurus [2].

Information retrieval thesaurus is a controlled vocabulary of domain terms that is created to improve the quality of information retrieval in that subject area. In other words, it is a collection of language units that make sense of the semantic relationship between them. Its goals are shown in the Figure 5.

Building an information retrieval thesaurus consists of several interrelated steps. The first stage is the formation of a dictionary. It is the initial set of keywords. This examines the array of the most informative documents for a given subject area.

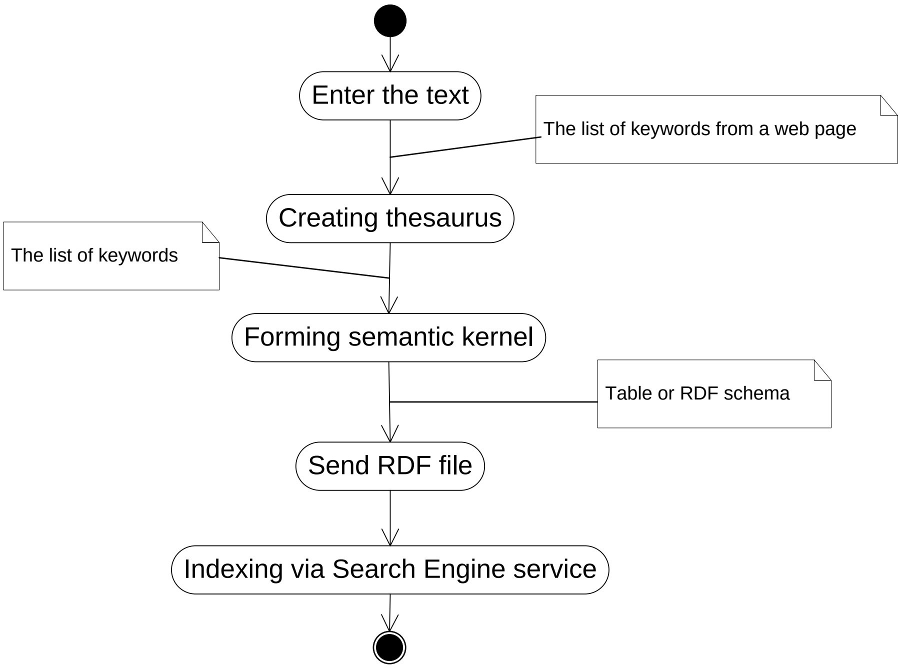


Figure 5: Solution schema

The second stage is the formation of multiple keywords. A set of keywords is formed from the dictionary. The selection of keywords analyses the informative nature of the word, which is determined based on the frequency of occurrence of the word, the role of the word in this subject area.

The process of choosing keywords is difficult to formalize. For example, a criterion such as frequency of occurrence cannot be absolute. If a word occurs very often in texts, it may indicate that it is too broad, or not well defined, that is, uninformative. If a keyword is rare, it may mean that it expresses a new concept and is thus informative.

The third stage is the formation of equivalence classes. Automatic information retrieval thesaurus is an integral part of automatic indexing of documents and queries.

An automatic thesaurus dictionary article typically has conditional equivalence relationships, subordinate links, and associative links. However, all these aspects were discussed in [2] fully. Thus, the algorithm is defined and represented in Figure 5.

The analysis of methods and tasks was provided – please, check Table 1. At the intersection of the row and column, the "+" symbol indicates the conformity of the method and functionality it solves.

The analysis of Table 1 shows us that there are not any appropriate methods for task solving.

Thus, we have regarded the alternative published here [2]. This is a special algorithm, which is based on some methods mentioned above.

4. Proposed IT solution

Using UML [9], the following use case diagram was designed (Figure 6).

According to the functional requirements done by a user (Figure 4), the software has two main users: a user and an administrator (developer), who shares the functionality.

The diagram illustrates the fact that a user having a product can compile the keyword dictionary (semantic kernel). Then, over time, the vocabulary is becoming more effective. This enables the buyer to find the product he is looking for faster because of the value for money.

Table 1
Overview of methods of problem solving

Method	Keyword rank	Normal form of keyword	Link analysis of keywords
TF-IDF	+	-	-
Lemmatization	-	+	-
Relevance feedback	+	-	-
Decision trees	-	-	-
N-grams	-	-	+
New approach [2]	+	+	+

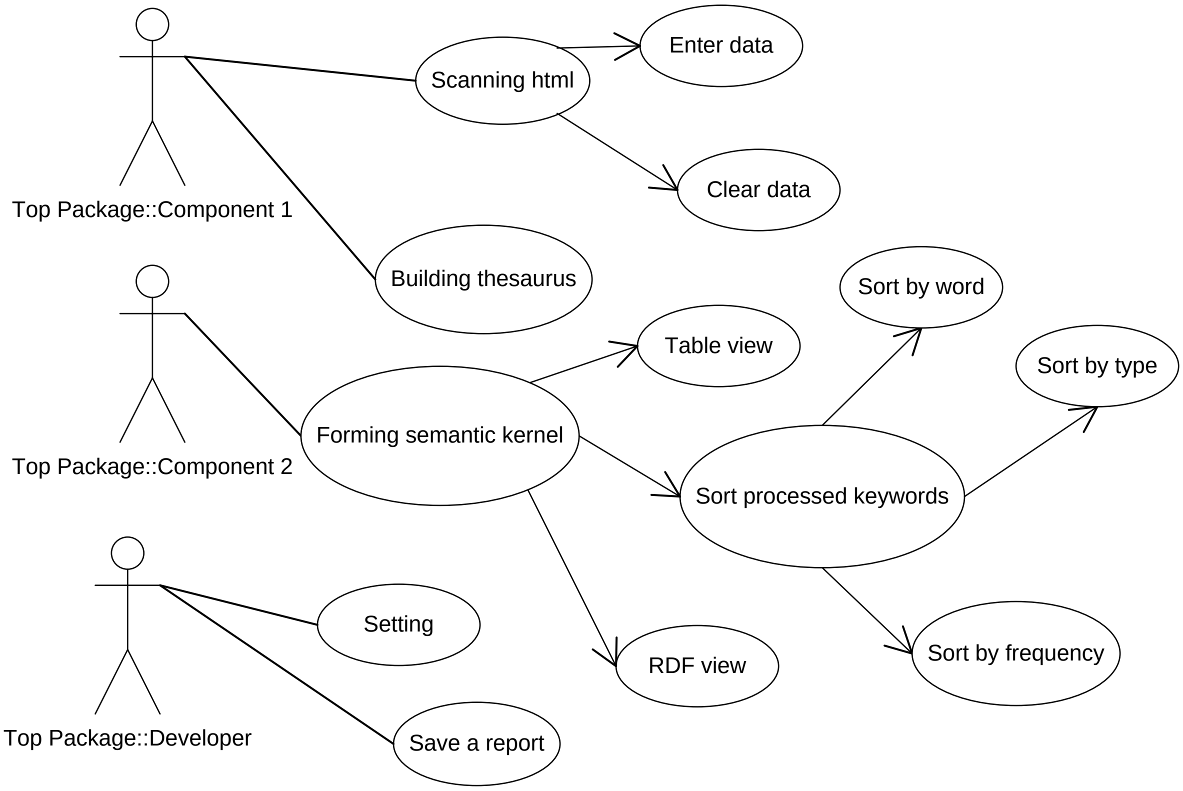


Figure 6: Use cases

The conceptual data model is shown in Figure 7 and reflects the structure of the projected keyword dictionary. The conceptual data model was built as a class model. It depicts logical structure of information that we have.

The algorithm of software work is presented in Figure 8. To see the software architecture the deployment diagram was prepared – Figure 9.

The main page of the developed software is shown in the Figure 10.

The core includes a set of keywords that is dynamically changing depending on market events over time. For convenience, use the block to order the keywords – Figure 10. In this block, a user selects the number of tags that need visualizing, as well as the depth of this analysis.

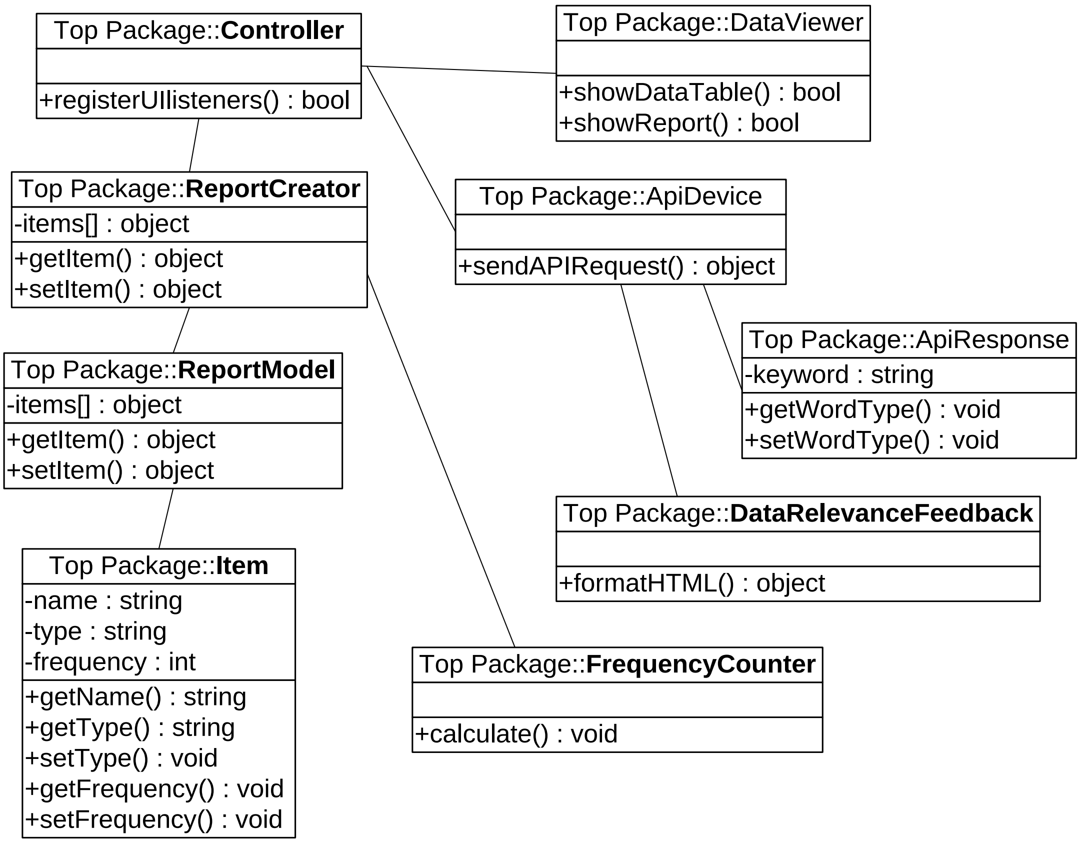


Figure 7: Class diagram

The constructed dictionary for a certain period is a table of keywords and calculated “frequency” score values. A number of keywords is displayed as RDF schema – Figure 10.

The resulting semantic kernel in the form of RDF document is used to promote this WEB resource in the search engine.

Finally, the software was developed as JavaScript application based on NodeJS framework [10].

5. Results

An English-language website on astrological knowledge and prediction was chosen to test the software components. Figure 11 shows a part of the text of the main page of such a site. According to the user's instructions, a thesaurus and a semantic core were built.

As an intermediate stage, a special thesaurus of keywords was formed according to the method [2]. Each keyword has a weight, which was interpreted as the number of rules that are set between words in the real text (Figure 11).

The obtained software was used to study the change in the semantic kernel of web project during 2017 - 2020. The project has been functioning on the Internet since 2009. Its subject is an astrological service for predicting future events.

According to the convention management system, the site has been changed at least three times. Each such alteration was accompanied by a revision in the main set of keywords (semantic kernel). Three copies of the site content were stored in the system, which allowed tracking the change of the semantic kernel (Figure 11). With the help of the Google Analytics service, you can get a relationship between the time when the kernel was changed and the number of visits to the site.

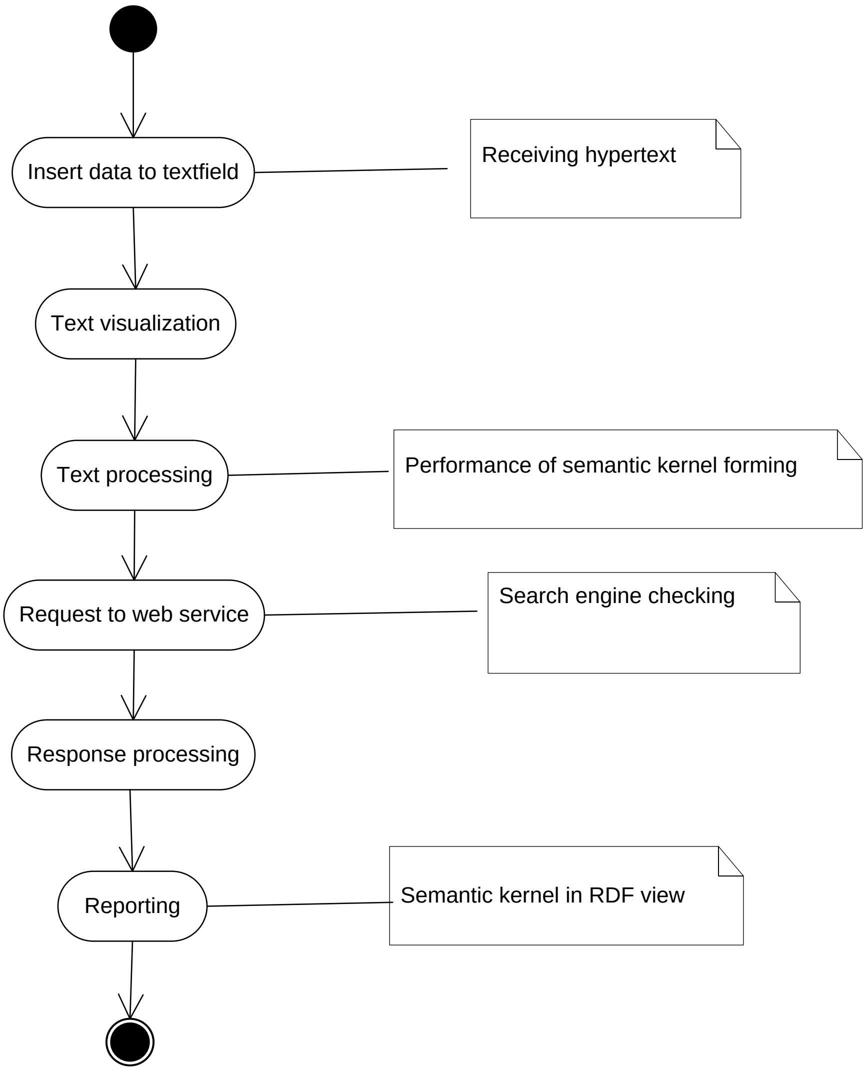


Figure 8: Activity diagram

It is possible to notice that each change of a kernel led to an increase in visits. But an interesting effect was also found, which was called kernel aging.

The fact is that the opinion of real users about astrology and so on is definitely changing over time. Therefore, the keywords (the kernel) that was embedded in the content of the site a year ago may no longer satisfy the user, that is, he is looking for other keywords on this topic that are not a part of this website, so the number of visits is decreasing. Nevertheless, changing the kernel corrects this effect.

The effect of kernel aging is very clear in web project – Figure 11. Here, the last restoration of the kernel took place in 2017. As a result, the site rolled from the second page in the search engine's responses to the tenth. This is also evidenced by web counter data from Google Analytics.

Thus, in the work by means of the developed software (Figure 10) an enthralling effect of kernel aging was revealed, which leads to a decrease in the efficiency of search engine optimization of the web resource.

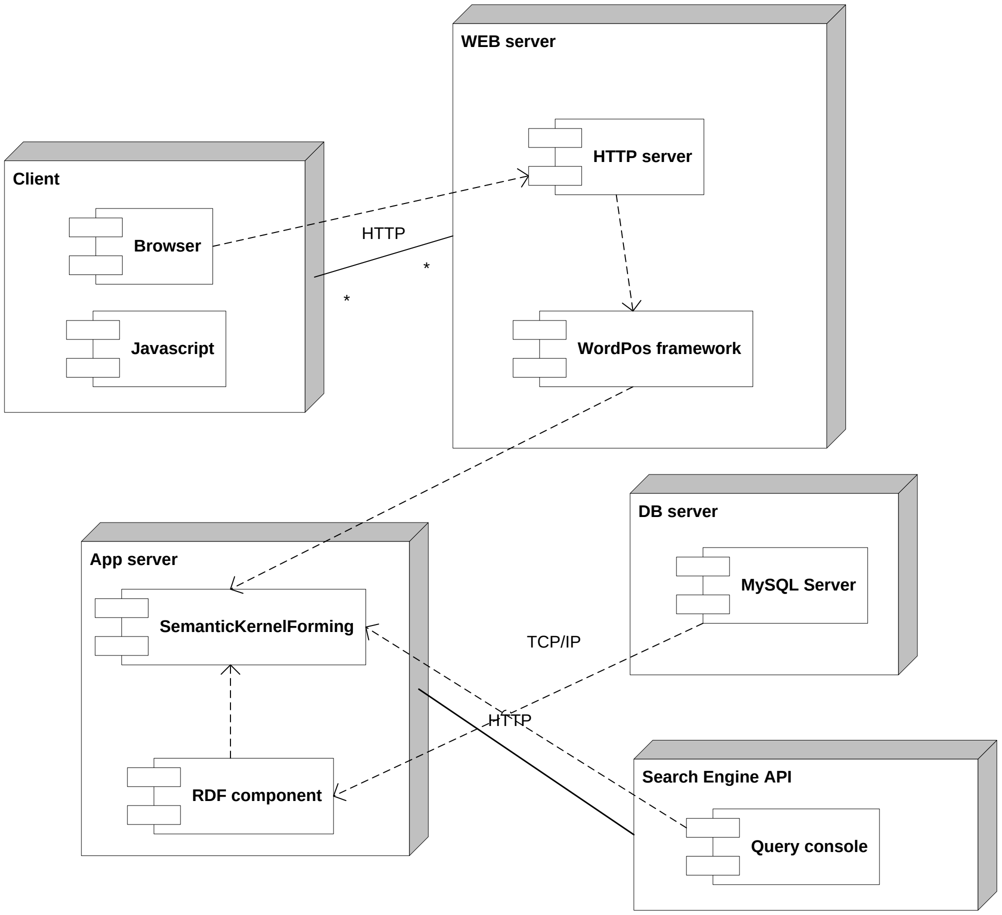
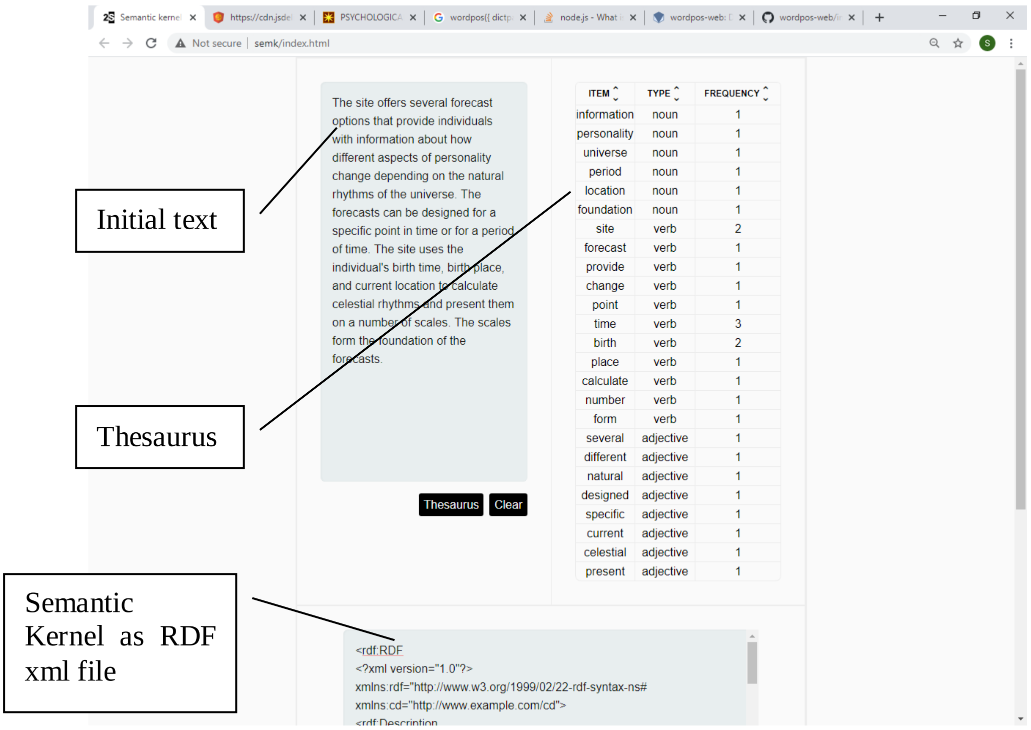


Figure 9: Deployment diagram

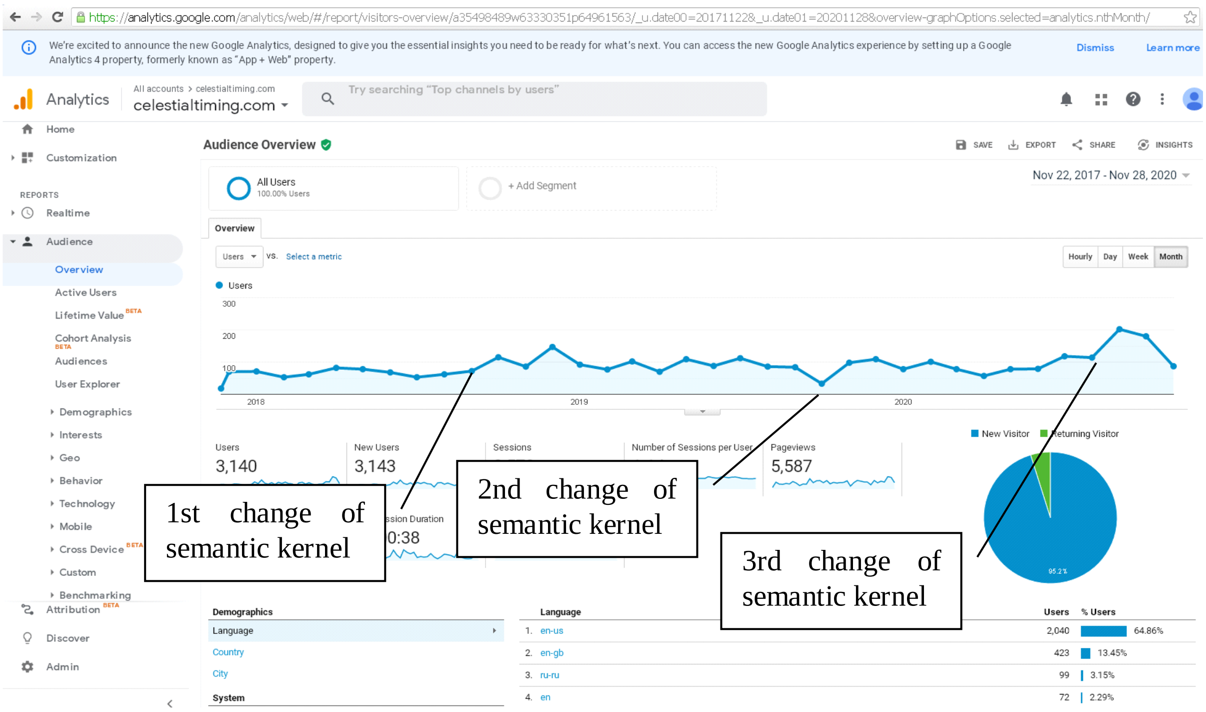


The site offers several forecast options that provide individuals with information about how different aspects of personality change depending on the natural rhythms of the universe. The forecasts can be designed for a specific point in time or for a period of time. The site uses the individual's birth time, birth place, and current location to calculate celestial rhythms and present them on a number of scales. The scales form the foundation of the forecasts.

ITEM	TYPE	FREQUENCY
information	noun	1
personality	noun	1
universe	noun	1
period	noun	1
location	noun	1
foundation	noun	1
site	verb	2
forecast	verb	1
provide	verb	1
change	verb	1
point	verb	1
time	verb	3
birth	verb	2
place	verb	1
calculate	verb	1
number	verb	1
form	verb	1
several	adjective	1
different	adjective	1
natural	adjective	1
designed	adjective	1
specific	adjective	1
current	adjective	1
celestial	adjective	1
present	adjective	1

```
<rdf:RDF
<?xml version="1.0"?>
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#
xmlns:cd="http://www.example.com/cd">
<rdf:Description
```

Figure 10: Main page



https://analytics.google.com/analytics/web/#/report/visitors-overview/a35498489w63330351p64961563/_u.date00=20171122&_u.date01=20201128&overview-graphOptions.selected=analytics.nthMonth/

We're excited to announce the new Google Analytics, designed to give you the essential insights you need to be ready for what's next. You can access the new Google Analytics experience by setting up a Google Analytics 4 property, formerly known as "App + Web" property.

Nov 22, 2017 - Nov 28, 2020

Users 3,140 · New Users 3,143 · Pageviews 5,587 · Session Duration 0:38

	Language	Users	% Users
1.	en-us	2,040	64.86%
2.	en-gb	423	13.45%
3.	ru-ru	99	3.15%
4.	en	72	2.29%

1st change of semantic kernel

2nd change of semantic kernel

3rd change of semantic kernel

Figure 11: Software testing

6. Conclusion

The performed scientific work allowed us to get some interesting results:

1. The study proposes a new approach to the implementation of the concept of the semantic web based on the concept of the semantic core. The implementation of this approach is based on the RDF schema. This allows implementing all four components of the Semantic Web without violating the principles of the modern Internet.
2. The implemented software was tested based on a real web project that has existed on the web since 2009. An interesting effect of the aging of the semantic core was revealed. The physical meaning of this effect was shown.
3. The proposed software can be easily integrated into existing web applications. Therefore, the goal of the future research is to accumulate factual material on changes in semantic kernels in various subject areas and to establish a relationship between them and real changes in the preferences of end users.

7. References

[1] J. Vishal, S. Mayank. Ontology based information retrieval in semantic web: a survey. International Journal of Information Technology and Computer Science, Vol.5, No.10 (2013) 62-69. doi: 10.5815/ijitcs.2013.10.06

[2] 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.

[3] S. Orekhov, H. Malyhon, I. Liutenko, T. Goncharenko, Using Internet News Flows as Marketing Data Component in: Proceesings 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] G. Lancaster, L. Massingham. Essentials of Marketing Management. 2nd. ed., Routledge, London, 2017. doi: 10.4324/9781315177014

[5] S. Godin. This is marketing. Portfolio/Penguin, USA, 2018

[6] F. Kotler, K. Keller. Marketing management. Pearson, USA, 2015

[7] F. Kotler F., G. Armstrong. Principles of marketing. Prentice Hall, USA, 2011

[8] I. Witten. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, USA, 2011

[9] B. Rumpe. Agile modeling with UML. Springer, Germany, 2017

[10] J. Lengstorf, K. Wald. Pro PHP and jQuery. APress, USA, 2016