Using semantic field model to create information search engines

V E Sachkov¹, D O Zhukov¹ and E G Andrianova¹

¹ MIREA - Russian Technological University, 78, Vernadsky ave., Moscow, 119454, Russia

E-mail: megawatto@mail.ru

Abstract. In the web infrastructure of information search, the use of semantic methods is considered to be a new round of the technology development. With the emergence of big data, a relevant issue is processing large amounts of data to extract valuable knowledge, especially for text files in natural language. Practice shows that traditional natural language search engines cannot always extract the necessary data from such data sets, as they do not take into account several subtle aspects of the language used in human speech. To solve this problem, the possibilities of using semantic search engines for text processing are being explored. This paper discusses the possible use of the semantic field model developed by the authors, to create a semantic search engine. Experiments have shown that using this model can improve the search accuracy. This model can be additionally used in creation of interactive dialogue systems.

1. Introduction

In the modern world of big data [1], the issue of processing large amounts of data to extract valuable knowledge is relevant. Text files in natural languages are one of the most important types of generated digital information. Text arrays published in the Internet contain significant amount of knowledge. Extracting this knowledge is possible subject to analyzing the materials of mass media, social networks, fora, blogs, news portals, etc. Practice shows that traditional natural language search engines cannot always extract the necessary data from such data sets, as they do not take into account several subtle aspects of the language used in human speech [2]. Today, to solve this problem, the possibilities of using semantic search engines for text processing are being explored, which allows for searching not by keywords or by the frequency of terms in documents, but by the semantic content of texts [3]. This means searching for similar documents in terms of meaning, identical subjects, and tonality. Therefore, we can conclude that the most relevant and sought-after area in text processing is the development of semantic search models, algorithms, and engines.

The basis of any semantic system for natural language processing, which would allow for finding and storing semantic relations between elements, is identification of semantic structures.

To date, there are many embodiments of such structures:

- Semantic network is a graph to present knowledge in the form of interconnected nodes and arcs [4].
- Conceptual graphs are a finite connected bipartite graph with nodes of two types, one type corresponding to concepts, and the other one, to their conceptual relations [5].
- Frames: a text data structure representing its stereotypical situation, an abstract pattern or a template for a situation, an object class, or a single object [6].
Formal ontologies: a system of concepts represented as a set of entities connected by various relations. Ontologies are used for formal specification of concepts and relations characterizing a particular area of knowledge.

The above structures describe well all the subtleties and abstract semantic connections of texts; however, their common drawback is the difficulty of representation of such structures in a formalized way for computer processing.

Another way to create formalized semantic structures is using the Semantic Web concept, where one of the popular models is the data presentation model – Resource Description Framework (RDF). However, the above model also has several disadvantages for use: a complex semantics of the query language; heterogeneity and incompatibility of ontologies; a low prevalence of the standard, etc.

It should be noted that another difficulty consists in the fact that most semantic systems are operated manually. Document bodies are marked up manually; there is a large share of the operator’s work in analyzing the collected information. Therefore, this area needs developing automation tools to reduce the human involvement in this process.

The semantic field model considered in this paper allows solving some of the above problems, being designed to create a convenient semantic structure for storing in a computing environment; to simplify operations with data; to facilitate the process of creating semantic search engines and systems for automation of the document bodies markup.

2. Semantic field model

One of the main problems in creating semantic word processing systems is the formalized presentation of semantic structures for storing the meaning of a text or document in a natural language, that would be convenient for storing and using in computer processing. To solve this problem, the semantic field was chosen as the basis for the semantic structure use.

Semantic field is an object uniting words of various parts of speech based on a common semantic attribute.

To connect keywords with dependent words, associativity of terms can be chosen as a semantic attribute. Associations make it possible to abstract from the direct meaning of the word, which allows replacing them with a set of other words. This has the opposite effect: based on a set of words (associations), we can restore the searched word [7]. This attribute allows creating a search query without knowing the keywords or terms of the subject area that the user does not understand; nevertheless, the user obtains the required result. In addition, this semantic attribute facilitates the automated construction of a semantic field.

Figure 1 shows the developed conceptual model of a semantic field. A text document in a natural language is transformed into a semantic structure consisting of a set of keywords or terms.

![Figure 1. Conceptual model of a semantic field.](image-url)
Each word has dependent lexical units (words or phrases) based on associative attributes (a semantic field). Such a structure makes it possible to convert the document semantic data into a formalized presentation that can be stored in computer memory.

The main purpose of the model is the automated creation of a semantic field and the extraction of associative context. Using the previously discussed features, we developed a semantic field model based on association properties.

Figure 2 shows an UML diagram of activity for the developed semantic field model. The model consists of two activity sections: creating a semantic field structure and extracting an associative context (upon the user’s request, a connected set of associations is returned).

![UML diagram of activity for the developed semantic field model.](image)

Let us consider the developed model in detail. The main steps of its implementation at the first stage are:

1. Creating or obtaining a body of documents in a natural language, defining the semantic field context in a specific subject area.
2. Text preparation: clearing it from punctuation marks, removing “stop” words, stemming.
3. Creating a dictionary matrix: creating a vector space where each vector of the text document consists of dictionary elements of unique words included in the body of documents.
4. Obtaining keywords: creating a set of words or n-gramms, based on which associations will be searched.
5. For each keyword, the frequency matrix of association words in the document body is calculated.
6. Creating a matrix of associations, — a document indicating the frequency of associative terms from the created association dictionary in the vector elements. After that, a formalized structure is generated, representing the semantic field.

The main activities at the second stage are:

1. Receiving a request to search for an associative context in the semantic field.
2. Data from the semantic field is extracted through calculations in the vector space based on distributive semantics.
3. Issuing a ranked response.

**3. Creating a Semantic Field**

The method of creating a semantic field is based on calculating the word frequency for syntagmatic and paradigmatic associations. It follows from the association properties that syntagmatic associations are in a window at a distance of 1 to 3 words from the keyword, whereas paradigmatic associations, in a window of 3 to 5 words. These properties allow for the automation of the process of obtaining the same
by splitting the text into n-gramms and by selecting terms with the highest frequency for a specific keyword.

As the basis of data processing in the practical implementation dealing with a semantic field model, we have chosen the neural network models: CBOW and Skip-gramm, as described by Thomas Milkolov [8]. These neural network models create two-dimensional representations of words, allowing to capture the frequency relations between words using statistical methods, which makes it possible to use simple algebraic operations with word vectors.

To create a semantic field, we used a body of specially prepared documents from the “RIA NEWS” news portal (https://ria.ru). The body consists of 250 thousand documents in a natural language from 13 main website sections for 3 years of news publications on the portal.

An example of trained neural networks using sliding windows of 3 to 5 n-gramms for storing the semantic field structure is given in Figure 3. When using the CBOW model, the following test words were sent to the network input: airplane, phone, 2010, USA, democrat, million, Putin, Foreign Ministry, science, Trump, Turkey. Using the T-SNE method, we reduced multi-dimensional vectors to two dimensions. Data having an implementation probability from 0.5 to 1 at the output is shown in Figure 3. Each word that is close to the test one has been assigned a specific color and a type of label display on the chart, whereas the test words were marked with text at their location in space.

Figure 3 shows that a specific group of words is concentrated, forming “clouds” around the keyword. The closer a word to the keyword, the higher its frequency in the window from 3 to 5 words. This allows automating text markup to search for associations by frequency without the user’s intervention.

When automating the process of creation of a semantic field with an associative context, there is a difficulty related to extracting the context due to semantic ambiguity. The reason is that the uniqueness of words in a particular subject area may be overlapped by their frequency in another subject area, thus issuing a wrong context in response to a search query. To eliminate the semantic ambiguity, it is possible to use semantic cores.

![Figure 3. Visualization of multi-dimensional vector space of the CBOW model reduced to two dimensions using the T-SNE method.](image)

The semantic core is the vector space of the semantic field, where the search is carried out for words and associations close in meaning.

The relevant semantic core is chosen using a modified naive Bayesian classifier that selects a semantic core matching the search query context.

4. Algorithm for Finding and Evaluating Syntagmatic and Paradigmatic Associations

The semantic field is used only to extract the semantic context according to paradigmatic and syntagmatic attributes; therefore, it makes sense to create specialized associative vector
spaces (AVS) [9] for storing the associative context associated with the corresponding natural language text or document, allowing to extract semantic context and to perform semantic search.

To find syntagmatic and paradigmatic associations that characterize the document text and are used for semantic search for semantically similar documents, an algorithm for evaluating the associations found has been developed.

The functional diagram of the algorithm is shown in Figure 4.

**Figure 4.** Functional diagram of the algorithm for Finding and Evaluating Syntagmatic and Paradigmatic Associations.

The main steps of the algorithm are:
1. Splitting the input documents into n-grams.
2. Finding all association vectors using the semantic core.
3. Evaluating the association vectors using the association evaluation module.
4. Generating the results. The algorithm result consists of two data sets:
   a. Summary: the n-gram used to obtain the set of associations (the associative context).
   b. Associations: the set of associations obtained, related to a particular n-gram.

The method for finding an associative context in the semantic core consists in comparing the input vector semantic proximity with the available vectors in the core dictionary. The semantic proximity of two vectors is calculated as the distance between such vectors, using a cosine similarity measure:

\[
similarity(A, x) = \arg \max_{x \in V} \left\{ \frac{\sum_{i=1}^{n} A_i x_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \cdot \sqrt{\sum_{i=1}^{n} x_i^2}} \right\}
\]

Where \(A_i\) is the input request (text, n-gramm, etc.), converted into a vector; \(V = \{x_0, x_1, \ldots, x_n\}\) is the set of semantic core dictionary vectors; and \(x_i\) is the word vector in the dictionary. Then, the set of word vectors forming an associative context is represented by the following condition:

\[
W = \{x \in V | similarity(A, x) \geq k_{\text{min}}\}
\]

Where \(W\) is the set of association vectors, and \(k_{\text{min}}\) is the threshold factor of the minimum cosine distance between the vectors, as required to filter the results obtained (the separation metrics). The complexity of implementing this logic lies in the fact that, according to formulas, the value of \(k_{\text{min}}\) shall be specified.

To determine the factor \(k_{\text{min}}\), we need to understand how many words we need to input to evaluate associations. Figure 5 shows the cosine measure of the first word distance (dark gray), depending on the input n-gramm of 3 to 10 words, as well as the dependence on the average mathematical expectation of the first 5 closest words (light gray).

The chart shows that, the larger the n-gramm size, the worse the output result; n-gramms of 3 to 5 words show the best results. The mathematical expectation of the 5 closest words provides approximately the same result, which is unsatisfactory for solving semantic search problems. From the
data shown in Figure 5, we can conclude that the best results are provided by the evaluation of 3 to 5 first words.

Figure 5. Functional diagram of an algorithm for finding and evaluating syntagmatic and paradigmatic associations.

Let us consider the histogram in Figure 6 that provides the cosine distances for the first 5 closest words. The histogram shows that the closest word having the greatest value in the distribution features the greatest accuracy and evaluation completeness. The maximum distribution peak falls on the value $k_{\text{min}} \approx 0.75$.

Figure 6. Functional diagram of an algorithm for finding and evaluating syntagmatic and paradigmatic associations.

After determining the average factor value, let us give an example of using the algorithm when analyzing a news document in Russian from the “RIA News” portal:

“БЕЙРУТ 10 авг. РИА Новости, Михаил Алаеддин. Самолеты ВВС международной коалиции во главе с США нанесли удары по жилым кварталам города Дейр-эз-Зор и по поселению Ар-Табанни, 19 мирных граждан погибли, сообщает сирийское национальное агентство САНА. В Ираке опровергли данные об ударе коалиции по шиитскому ополчению в Сирии. Удары были нанесены по северным окраинам Ар-Табанни, какие именно кварталы города были атакованы, агентство не уточняет. В конце июля самолеты коалиции разбомбили больницу Аиша в городе Абу-Кемаль в провинции Дейр-эз-Зор, тогда погибли шесть человек и десять получили ранения. Россия неоднократно подчеркивала, что действия в Сирии США и международной коалиции не легитимны, так как проводятся без согласия и координации с официальным Дамаском».

In English:

“BEIRUT, Aug. 10. RIA News, Mikhail Alaeddin. Air force aircraft of the International Coalition led by the USA attacked residential areas of Deir ez-Zor and the town of At-Tabanni; 19 civilians were killed, as reported by Syrian national agency SANA. Iraq refutes information on the coalition attack on Shiite militia in Syria. The attacks were inflicted on the northern outskirts of At-Tabanni; the Agency...”
does not specify which town districts were attacked. In late July, Coalition aircraft bombed the Aisha Hospital in the town of Abu Kemal in the province of Deir ez-Zor; then, six people were killed, and ten people were injured. Russia has repeatedly stated that the actions of the USA and the International Coalition in Syria were not legitimate, as they were carried out without consent and coordination with the official Damascus.”

A partial result of processing this Russian-language publication is provided in Table 1 (words in Table 1 are shown in the processed short (stemmed) form).

| No | Annotation | Cosine distance | Associative context |
|----|------------|-----------------|---------------------|
| 1  | шиитск ополчен сир удар нанес | 0.816580 | авиадар, массирова, воздух, хусит, курд, иракск, суннит, террорист, повстанц, нараста |
| 2  | нанесл удар жил | 0.810924 | нанесен, авиадар, воздух, нанос, танк, слезоточив, бомб, уничтож, разруш, авиа |
| 3  | сша нанесл удар жил квартал | 0.769283 | нанесен, воздух, авиадар, нанос, баллистическ, ракет, тип, несанкционирован, слезоточив, бомб |
| 4  | конц самолет коалиц разбомб больниц | 0.758578 | укрыва, оттуд, гранат, унесш, банд, браун, боин, бойн, ливнев, затонул, бригад |
| 5  | Самолет ввс международн | 0.749910 | сбит, а321, авиалайнер, бомбардировщик, малайзийск, Boeing, су, сбивш, пилот, боин |

In Table 1, associative context is collected into a single set and converted to a frequency vector, similar to a semantic field dictionary, where each such vector links associations and the current document. A set of such vectors, linking each document in a document body by an associative context, forms an associative vector space (AVS).

5. Using a Semantic Field Model in Associative-Semantic Search

It is convenient to use a semantic field model, together with an association evaluation algorithm, for arranging an associative semantic search, where the document similarity is compared by AVS vectors [9]. “Earth movers distance” (EMD) is used as a metric for comparing two documents [10]. The EMD computation is based on the solution of a transport problem, where the distance matrix is calculated, and the results are issued based on the calculated parameters ranking of this metrics.

The semantic search in the associative vector space was tested using the methods of search, selection, and ranking of thematically similar documents based on the collected document bodies, while using the same document body as for training the semantic field model. As an example, the “Aviation Accidents” topic was selected in the form of a small collection of documents.

Figure 7. Functional diagram of an algorithm for finding and evaluating syntagmatic and paradigmatic associations.
Next, two documents with numbers 119 and 121 were searched. The search result interpretation is given in Table 2.

### Table 2. Decoding Search Results.

| No | EMD   | Text                                                                 |
|----|-------|----------------------------------------------------------------------|
| 1  | 67.171673 | москв 16 — прайм представител межгосударствен авиацион комитет мак принима участ работ коммис миноборон рф расследован катастроф самолет 154 произошл соч 25 декабр прош сообщ комитет расследован дан катастроф привлеч специалист лаборатори исследовательск баз науч техническ центр мак отмеча ко... |
| 2  | 78.862919 | москв 15 — пассажирск самолет франкфурт — москв немецк авиакомпани Lufthans вечер воскресень соверш аварий посадк аэропорт домодедов российск столиц перегрев любов стекл сообщ источник экстерн служб предваритель причин аварий посадк — перегр любов стекл — собеседник 22 30 мск московск аэропорт ... |
| 3  | 90.808066 | краснодар 14 — татья кузнецов пассажирск самолет Boeing летевш томск суббот вечер попыт ослеп лазер анап сообщ воскресень представител южн транспортн прокуратур инцидент произошел 21 30 мск взлет Boeing рейс анап — томск освещ лазерн луч зелен цвет ослеплен член экипаж полет повлия — рассказ собе... |
| 4  | 91.159623 | петербург 1 — вертолет ми 8 борт могл наход высокопоставлен чиновник ночь воскресень потерпел крушен мурманск област борт наход 18 дво выж настой обнаружен тел погибш опозна сообщ источник правоохранитель орган регион тел поднят водолаз поверхн опозна заместител губернатор мурманск област сред — со... |
| 5  | 93.339509 | рост дон 20 — аэропорт ростов дон утр суббот произошл крушен пассажирск Boeing официалн закр 08 00 утр понедельник сообщ представител аэропорт ростов дон официалн закр 08 00 утр 21 март — поясн лайдейниц добав связ ка задерж восстановлен взлетн посадочн полос днем ... |

The result analysis is shown in Table 3, indicating:

- Matches of unique words from the selection: the number of matches of unique words in the document found with words from the search selection, as a percentage value.
- The use of unique words from the selection: the number of matching unique words from the total number of the search selection.
- Association matches in the documents: the value characterizing the similarity of the AVS vector in the document found with the AVS search selection.

### Table 3. Analysis of search decryption results.

| Document Number | EMD   | Match unique words from a selection in a document % | Total words in a document | Used unique words from the sample % | The number of associations in the document | Matching associations in document and sample % |
|-----------------|-------|-----------------------------------------------|--------------------------|-----------------------------------|------------------------------------------|-----------------------------------------------|
| 119             | 67.17 | 100.0%                                        | 57                       | 54.3%                             | 69                                       | 100.0%                                        |
| 121             | 78.86 | 100.0%                                        | 51                       | 48.6%                             | 63                                       | 13.0%                                         |
| 485             | 86.86 | 11.7%                                         | 429                      | 47.6%                             | 598                                      | 68.1%                                         |
| 17173           | 90.81 | 19.4%                                         | 129                      | 23.8%                             | 262                                      | 29.0%                                         |
| 19281           | 91.16 | 11.6%                                         | 328                      | 36.2%                             | 480                                      | 40.6%                                         |
Let us consider the first two lines. Document 119 contains all the words from the search selection (100%), as well as document 121. However, the AVS vector ratio for 121 is 100%, and for 119, 13%; they also have the smallest EMD distance. Further study revealed that the matches of unique words in the document and in the search selection vary on average from 10% to 20%, whereas the matches of AVS vectors, from 10% to 60%; the use of words from the selection, from 15% to 55%. Such a difference in values suggests that matches of words or AVS vectors do not matter; the main parameter is EMD, as a document ranked, e.g., at the 100th position in terms of EMD, may have a 60% word match and a 60% AVS match, while being completely irrelevant to the search topic.

In more details, see about the associative semantic search in [9].

This approach has well proven itself in creating interactive dialogue systems. The use of interactive dialogue systems greatly facilitates the interaction between the computer and the user due to the fact that the communication between them is carried out in a natural language. The user of such a system does not have to possess some special skills and knowledge. In addition, if the interactive dialogue system is capable of a meaningful dialogue, this increases the user's convenience and trust in the system [11].

6. Conclusion
The use of the semantic field model facilitates the semantic structure formalization for computers by converting semantic relations into a set of vectors based on associative attributes. The properties of syntagmatic and paradigmatic associations allow creating systems for building up of the represented model based on the word frequency. This makes it possible to use the existing solutions and to simplify the work on marking up the document body used for training. A model with an association evaluation algorithm allows getting a connected associative context with a natural language document, which forms associative vector spaces. Comparing vectors of associative vector space based on the EMD metrics allows finding semantically similar documents using associative attributes.

References
[1] Laney D 2017 3D data management: controlling data volume, velocity, and variety International Transaction of Electrical and Computer Engineers System 4 (1) 14-25
[2] Molchanov K V 2019 The dialectic study of the English language in the light of the work "Science of Logic" by Hegel: new dialectic cognition, dialectical programming, and the processes of thinking Russian Technological Journal 7 (2) 74-82 (in Russian)
[3] Lobanovsky Y I 2018 General classifier of system problems - I: analysis, synthesis, validation and verification Russian Technological Journal 6 (4) 5-12 (in Russian)
[4] Quillian M R 1968 Semantic Memory Semantic Information Processing (Cambridge: MIT Press) 227—70.
[5] Sowa J F 1979 Semantics of conceptual graphs Proceedings of the 17th Annual Meeting of the Association for Computational Linguistics 39-44
[6] Minsky M 1974 A framework for representing knowledge The Psychology of Computer Vision
[7] Rubinshtein S L 1997 Izbrannyye filosofsko-psikhologicheskiye trudy. Osnovy ontologii, logiki i psikhologii (Moscow: Nauka) 463 (in Russian)
[8] Mikolov T, Chen K, Corrado G and Dean J 2013 Efficient Estimation of Word Representations in vector space In Proceedings of Workshop at ICLR
[9] Sachkov V E, Zhukov D O, Korablin Y P, Raev V K and Akimov D A 2018 Organization of information system for semantic search based on associative vector space Communications in Computer and Information Science 1140 70–8
[10] Rubner Y, Tomasi C and Guibas L J 1998 A metric for distributions with applications to image databases IEEE International Conference on Computer Vision 59-66
[11] Sachkov V E 2018 The use of associative semantic preprocessor in interactive natural language dialogue systems Proceedings of the Institute of system programming of the RAS 30 (4) 195-208 (in Russian)