DOI: https://doi.org/10.15407/usim.2018.01.071

UDK 004.89

M. SAVCHENKO, Student (Masters), zitros.lab@gmail.com

O. KRIACHOK, Candidate of Engineering's Sciences, Associate Professor, alexandrkriachok@gmail.com

National Technical University of Ukraine ' Igor Sikorsky Kyiv Polytechnic Institute', Kyiv, boulevard of Victory 37, corps 5, 03056.

AUTOMATIC GENERATION OF SEMANTIC KNOWLEDGE NETWORKS FROM AN UNSTRUCTURED TEXT

A method and an algorithm for the semantic knowledge network automated construction created from the most informative concepts in the electronic texts are proposed. An analysis and comparison of existing methods with their software implementations for information research in electronic texts are presented. The results of BBC news article analysis using the proposed method are given.

***Keywords**: building semantic networks, knowledge extraction, knowledge models, natural language processing*

Introduction

In today's information environment, through the huge amount of unstructured text information, there is a need in the search, seizure, formalizing and processing of the most essential knowledge laid down by the authors in the texts. Such knowledge may be hidden in the concepts presented in the document, and the characteristic relationship between those concepts.

Considering the large number of texts, such as news articles or scientific publications, one can notice that each text has a certain, sense-unique characteristic only for the given text. Only after reading the text, one can briefly describe it, heading to highlight the most important concepts in it and combine their logical relations with other known concepts. For the rest of this document we will call every single text corpus as a document, a single word in it as a term and a word or group of words that represent a specific entity as a concept.

When it comes to a huge data set, Big Data or massive text corpus, just a single person can't read through it quickly, taking all the important information from there. There is a need to structure the knowledge in texts and present texts in a structured form that can be quickly analyzed. There are various methods of search and evaluation of the relevant information in the document on which one is about to make a decision using an in-depth analysis of certain parts. Additional statistical estimation methods (weighing) and the identification of the most relevant terms in the particular document include the following: [1].

Ranging the terms t_i in the document D_j by the number of a particular concept occurrences $t_i \in D_j$ — TF, which stands for Term Frequency. It is statistically investigated that if the document describes the specific area of knowledge, the most typical concepts in a particular document are repeated relatively a lot of times.

Ranging the terms t_i in the document D_j by the number of occurrences of a specific concept and inversely related to the total number of all other documents $D_j^{'} \forall j \neq i$, $t_i \in D_j$–TF-IDF, which stands for Term Frequency — Inverse Document

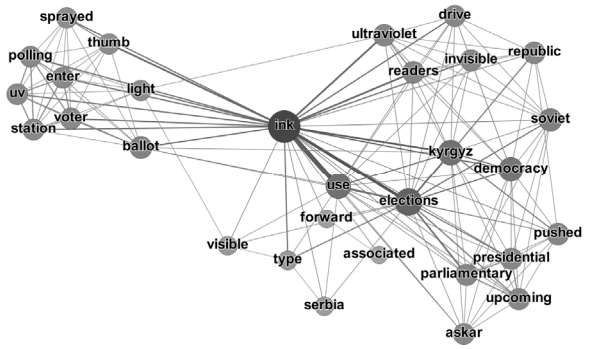

Fig. 1. Analysis of entities in the article using the matrix method

Frequency. In contrast to the TF, TF-IDF allows us to dramatically reduce the rank of terms found in almost all the documents, such as, for example, articles, prepositions, and other insignificant commonly used words.

Other methods, such as, for example, Okapi BM25, sigma method, etc.

The above methods allow us to highlight the most relevant concepts for a specific text. Using this information, when the relations between concepts are established, it can be dropped in case of these concepts a little weight hence giving priority to the concepts with the highest score.

To examine the documents and obtain the formalized representation of unstructured information that they contain, it is proposed to construct a semantic network of concepts and relationships between them. There are many approaches to the construction of semantic networks, and in this article only the most common ones are described.

Matrix method. To identify the most expressed relationship between two terms $\{t_i, t_j \mid i \neq j\}$, we can use the matrix method. In this method, at first, all terms are weighted by one of weighting methods (for example, TF-IDF), and each term t_i is assigned its unique corresponding weight p_i. After selecting n the most highly expressed set of terms, the adjacency matrix A is constructed and filled with zeros. Next, the document is divided into groups g_k. Each group may be represented, for example, as one individual sentence.

The adjacency matrix for the matrix method completes as follows. For each group g_k there is a pair of terms $\{t_i, t_j \mid i \neq j,\ t_i \in g_k,\ t_j \in g_k\}$. The number of groups into which a pair of these terms are carried, $\{t_i, t_j \mid i \neq j\}$ is added to the adjacency matrix A. In this case, we can also consider the weight of such pair of terms $\{t_i, t_j \mid i \neq j\}$ and use it as a coefficient for the resulting value in the adjacency matrix A.

To demonstrate the results, this article will use the text corpus of approximately 2500 BBC News articles in English, presented in a separate text document containing the title and full text of articles. Articles are divided into 5 main categories: business, politics, sport, entertainment and technology. Around 500 documents match each of the categories. For in-depth analysis we will use the article «Ink helps drive democracy in Asia», published February 19, 2005 on the website BBC News [2].

An algorithm for constructing semantic network using the matrix method [6]:

Take an evaluation (weighting) of each term in a given text in relation to all other terms in the text body by using one of the above methods (we will use TF-IDF).

Create a list of unique terms that are graded in descending order of their weight in the given text. 30 the most weighted terms are selected. In automatic mode, the analysis can be selected, for example, 3% of the most highly ranked terms.

Adjacency matrix is constructed by the algorithm described above.

The result of building is listed in the file with the extension *.csv and visualizing via Gephi visualization tool. The resulting graph is shown in Figure 1.

The above graph visualizes terms which have a connection with other terms in the text. The more text has the direct links between the two terms, the thicker is the connection between these terms on a graph. The size of a concept on a graph is also proportional to its weight of the weighting method.

It may be noted that concepts are mainly organized in separate clusters. For example, among other clusters, it can be noted that the most expressed terms are «voter enter uv station polling». Undoubtedly, this cluster expresses the most relevant part of the knowledge of the text, as compared to all other texts. Among the other most expressed clusters, such as «upcoming presidential parliamentary elections» or «Kyrgyz elections use

ultraviolet ink», you can get a clear idea of the nature and content of the text.

Matrix method includes the most well-interconnected terms, but one of the major disadvantages of this method can be regarded as the relative complexity of the allocating clusters which are the most expressed in the graph. It is not hard to do, looking at the graph visually with a small number of terms, but as the number of connections grows, it becomes very difficult to analyze and find such clusters, even using the software. In addition, for the resulting graph it is almost impossible to include such terms that have relatively low weight according to the algorithm of entities evaluation. This plays a key role for the data crawler.

Horizontal visibility graph application. Another interesting method for constructing semantic network of terms is the method proposed by Lande D.V., which combines features of the graph with the horizontal visibility evaluation methods in terms of a single text [3]. This method can be applied not only to build a network of basic concepts of terminology in the text, but also to build a semantic network as a whole.

An algorithm for constructing a semantic network for this method is as follows:

Text entities, similarly to previous method are evaluated (weighed) against relation to all other terms in the text body with TF-IDF method.

The algorithm for constructing the horizontal visibility graph applies to the values of the term weights. The horizontal axis is taken the term position in the text, and the vertical as a term weight. Before constructing the actual result, some procedures like stemming and rejection of the terms listed in the list of stop words are performed.

Figure 2 shows the principle of the visibility graph construction with the horizontal normalized TF-IDF values of evaluation. After term weights are put on the horizontal axis, for each term t_i in the document D a «horizontal search» applies to the corresponding document using a horizontal search algorithm [4]. Thus, two terms t_i and t_j are compared and, a bond is formed there between. The weight of this connection in the graph may be proportional to a predetermined «farsightedness» (horizon). In other words, terms that are adjacent

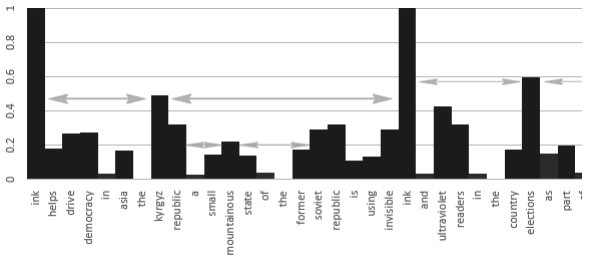

Fig. 2. The principle of the horizontal visibility graph construction

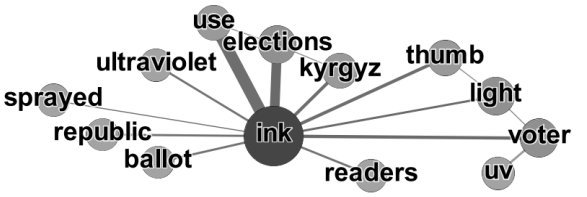

Fig. 3. The article analysis using the horizontal visibility graph algorithm

in the document form a strong bond, thereby form a strong relation between concepts.

One of the possible semantic network option construction for a given text using horizontal visibility graph is shown in Figure 3. This graph has been built with a threshold value of 0.25, which means that only those terms whose relative weight is greater than 25% get to the final result. The visibility horizon was set to 20 words, and the weight of a definite connection between the two periods was measured by the total amount of a linear distance between two terms.

Building a semantic network based on the approach of the horizontal visibility graph construction has an advantage in the formation of stable relations between concepts, as it allows us to explicitly highlight concepts and existing links between them through other concepts. But in this method, without modifications, the same disadvantages apply as in the previous one, matrix method: is difficult to determine the logical order of relationships between entities and, in addition, as in the matrix method, some medium or low valued words can be missed.

InterSystems iKnow technology. InterSystems Corporation is developing its own proprietary algorithms for in-depth analysis of the texts that also

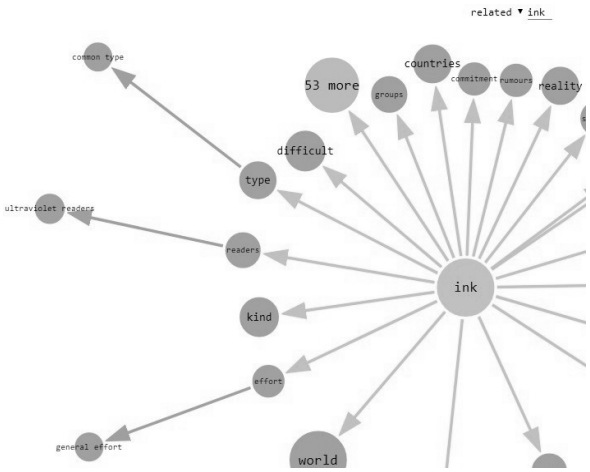

Fig . 4. The article analysis using iKnow technology with iKnow Entity Browser

can be used for the construction of semantic networks of words [7]. A set of tools for application analysis of texts, included in the InterSystems iKnow corporate product, allows to identify concepts in the text, the relation of similarity and connecting relation between these concepts. Any solutions can be built based on these tools for «structuring» unstructured information, such as, for example, a solution for semantic analysis of sentences, revealing modern trends through analysis of the news and so on.

The algorithm by which InterSystems iKnow finds the relationship between the concepts in the text, as well as similar concepts is mainly based on the use of the stable structures and words from natural language [8]. For example, some of the concepts in the text are resilient — they are changing rapidly, and there are too many to remember to find out the role of each concept in the text. But there are concepts which can be considered as relatively stable between different ages, for example: «no», «replace», «performing», «it is», «use», «used by», «stored in the» and so on. InterSystems iKnow recognizes such patterns for the specified language in the text and distinguishes separate concepts ratio there between from other minor parts of speech. Additional metrics are also computed for concepts, like the total numbers of concepts, their spread (the average distance between same concepts in the text), relevance and the total score, which is combined from the previous metrics by the formula.

To demonstrate the algorithm in action, as an example, let's take the sentence «clever cat eats cheese and breathes on a mouse burrows». iKnow technology will first consider the text as a set of sustainable language constructs, i.e. «__ __ eats and breathes on a __». After that, instead of «__» the other concepts are considered, whose presence there is almost guaranteed. The concepts also differ in terms of similarity, such that «smart cat» is similar to the concept of «cat» and «mouse holes» through «holes».

Thus, the article is analyzed and such concepts as, for example, «readers» and similar «ultraviolet readers», «effort» and «general effort» are identified. Figure 4 depicts a graph — word semantic network constructed on the basis of the article for analysis by InterSystems iKnow technology and visualized using the visualization tool iKnow Entity Browser. Note that the main concept of the graph is the «ink» – «Ink», from which arrows show which concepts it is related in the ink. The concept of the circle corresponds to the size of its assessment, which holds iKnow. The arrows coming out of these concepts reflect the «similar» concepts.

Thus, using iKnow toolkit can not only identify the most relevant concepts for this text, but also relationships between them and the nature of that relationships (ratio for denying the similarities).

Importantly, InterSystems iKnow concept is built mainly for working with entities, in fact, that basically requires a modern market. After gathering the necessary data about a particular concept using iKnow tools further expert should independently find all relevant entities in the text and to analyze individual sentences to update the basic knowledge that has been assigned to him. In addition, the technology is closed and is supplied with the main product of InterSystems' — DBMS Cachй (since 2018 — IRIS platform).

The complex method of constructing a semantic network. Taking into account all the advantages of the above-described methods for building semantic web (knowledge network) for a single text, the complex approach has been developed

and investigated that includes all the benefits of the methods described above and adds its own.

This method combines algorithms for evaluation of words with algorithms Part of Speech Tagging (POS Tagging, identification of parts of speech) that allows you to find the concepts with their characteristics in the text and to build relationships between them based solely on information obtained from the text and a small number of basic rules for a single language.

Also, this method is based on constructing a full semantic network (knowledge network) that contains the logical connection between the called and the concepts, as opposed to the simple binary «yes-no» relationships.

A simplified algorithm for constructing a semantic network using the complex method looks as follows:

Without a change in the original text, the identification of parts of speech (POS Tagging) is applied. It is important not to spend Stemming or normalization of terms at this stage, as for the identification of parts of speech the semantic meaning of the original text is important, as well as the register of symbols, punctuation marks, etc. As a result, each period is recorded in the pair identified with his part of the language.

Each term in the text gets its weight assigned, reduced to the normal form (lowercase Stemming for complex languages).

On the basis of the information received, concepts are determined in this text and, based on a language rules, relationships between concepts are added.

A semantic network is built

Steps 1–2 are presented in Figure 5. The timing marks are shown in a normalized form. The blue columns in the picture identify the concepts, namely, nouns and their corresponding adjectives (ink, Kyrgyz republic). Red bars represent verb (helps, is using). For convenience, let's call them functional terms. Green represents the auxiliary parts of speech (the, of, in, to).

Figure 5 shows that functional terms are almost always between two non-functional terms, i.e. concepts. Thanks to the definition of the parts of

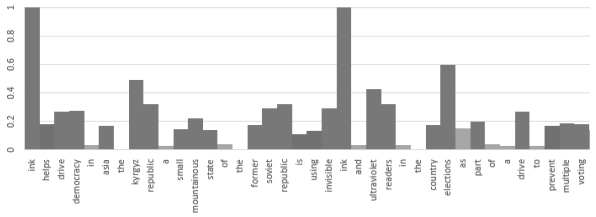

Fig. 5. Evaluated words and identified parts of speech

speech, it is possible to build the relationships between concepts from the given information. For example, on Figure 5 it is clearly observed two such relationships: «ink helps drive democracy» and «former soviet republic is using invisible ink». Thus, the algorithm picks three concepts: «(invisible) ink», «drive democracy», «(former soviet) republic», as well as the relationship between the two functional concepts of «helps» and «is using».

This idea of functional terms allocation is a few similar to the idea used in the InterSystems iKnow technology, but it is not identical. In contrast with InterSystems iKnow, complex method uses POS Tagging for identification of such terms but not pre-defined dictionaries. In turn, POS Tagging algorithm implementation may vary, including specially modified for a particular task, or one in which the neural network applied, which can greatly improve the accuracy of output results, even in the case of adding new words to the language or in case of mistakes in the text.

Algorithms for identifying parts of speech can now correctly recognize part of speeches with more than 97% accuracy [5], which is an acceptable value for constructing semantic network based on them.

To review and search for «neighboring» concepts and relationships, like in method with horizontal visibility graph, only those terms are taken which relative score is not less than a certain value. To construct the graph shown in Figure 6, the value of 0.25 or 25% was used. Increasing this value will add to the graph less relevant concepts, but also, in turn, increase the number of explicit relations in it and build logic circuits, which may lead to the discovery of new logical sequences in the text.

Unlike horizontal search using horizontal visibility graph, the complex method provides a com-

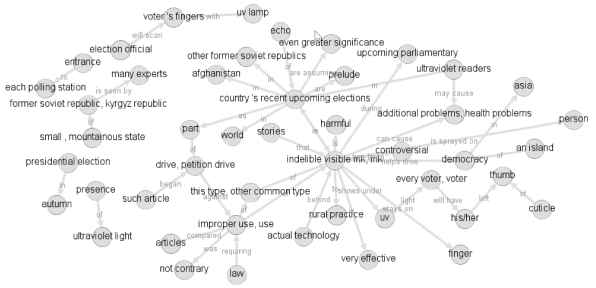

Fig. 6. A semantic network constructed using the complex method

Fig. 7. A semantic network built from another BBC article

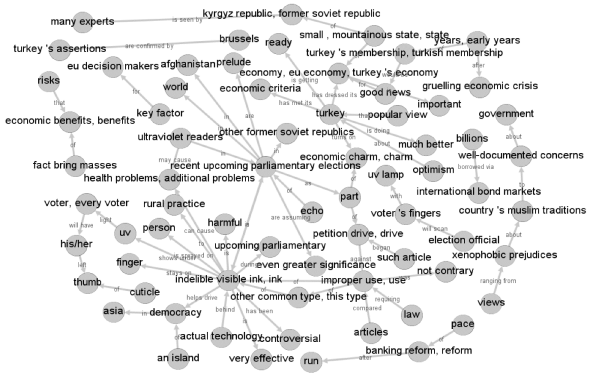

Fig. 8. A combined semantic network built from the both articles

prehensive way of breadth-first search (or searches extension of scope), starting from the term t_i and moving in a both sides. An important advantage of this approach is that the graph counts also those entities whose weight is relatively low, but only those that have a certain relationship with a highly weighted concept through the functional word.

For example, in the bar chart shown in Figure 5, the concept «drive democracy» has a rating less than 25% but, as is evident from Figure 6, it is present in the graph, because it is associated with more highly regarded concept «ink».

As a result of the depth-first search of concepts, starting from functional terms, a natural relationship between the concepts is formed in the semantic network, which does not distort the meaning of the original proposal. Moreover, the identical concepts are combined with many other concepts through a variety of relationships, it creates the possibility of a new way to treat the knowledge that has been invested in the text, without further distortion.

Figure 6 shows the results of complex text analysis method based on a BBC article «Ink helps drive democracy in Asia» [2]. Note that the obtained knowledge graph wasn't manually modified (except for the visual arrangement of the elements on the graph, which is performed using Gephi visualization tool).That is, it was built in a fully automatic way.

To provide a broader example of the developed complex knowledge graph building algorithm, we built knowledge graph of another article, «Turkey turns on the economic charm», published by BBC in approximately the same time frame comparing to «Ink helps drive democracy in Asia». The second knowledge graph is presented on Figure 7.

Moreover, to demonstrate that the retrieved data is accumulated and properly related using the same algorithm on multiple articles, we built a single knowledge graph from the two articles mentioned above, both published in BBC. The aggregated result is presented on Figure 8.

Joining multiple graphs into a single graph, which is presented on Figure 8 can be improved in future to consider filtering of even less relevant entities. For example, a set of texts may have many words in common. These common words can be considered when weighting concepts, by modifying TF-IDF metric with new parameters and hence cleaning the knowledge graph.

The algorithm, its implementation and the source code are published online [10], where one can find detailed instructions of how to reproduce the presented result and use this algorithm to generate semantic knowledge networks for new texts.

Conclusion. Complex method of the semantic networks combines construction all the advantages of the methods described in this article and does not inherit their drawbacks. With the help of

complex method, one can build semantic networks (knowledge graphs) from any number of texts in a fully automatic mode without the need of the system experts. In the results of data extraction from the basic text, the most relevant information presented as a graph knowledge can be used in the future, e.g., for the development of automatic intelligent analysis of any text data.

Designed information extraction algorithm, which considers only the most relevant information in the given texts is flexible and easily extendable. With just a few basic rules for the given language in the text, on average complex algorithm covers more than 50% of all entities in the English text. Within each rule, this percentage increases and could theoretically reach 100%, which is also reflected in the success of parts of speech recognition algorithms. By developing a set of rules for a particular language, it can be widely applied to any texts, and is not limited to only technical literature, but even to the individual texts written in a particular style of information representation.

REFERENCES

1. *Savchenko M.M.*, *Kriachok O.S.* Using models of knowledge for the analysis of unstructured text. Modern problems of scientific support for Energy. Proc. XV Int. scientific-practical conf. of graduate students, undergraduates and students. Kiev, Apr. 25–28, 2017, Igor Sikorsky KPI. 2. P. 121.
2. *Mikosz D.* Ink helps drive democracy in Asia. David Mikosz. 2005. Resource Access: http://news.bbc.co.uk/2/hi/technology/4276125.stm.
3. *Lande D.V.* Building of Networks of Natural Hierarchies of Terms Based on Analysis of Texts Corpora. E-preprint ArXiv 1405.6068.
4. *Lande D.V.*, Snarskii *A.A.*, *Yagunova E.V.*, *Pronoza E.V.* The Use of Horizontal Visibility Graphs to Identify the Words that Define the Informational Structure of a Text. 12th Mexican International Conference on Artificial Intelligence, 2013. P. 209–215.
5. *Toutanova K.*, *Klein D.*, *Manning C.D.*, *Singer Y.* Feature-rich part-of-speech tagging with a cyclic dependency network. In: NAACL 3. (2003), P. 252–258.
6. *Lande D.V.*, Snarskii A.A., Bezsudnov I.V. Internetika. Navigation in complex networks. Moscow: "LIBROKOM" Book House, 2009. 264 p.
7. *Van Hyfte*, *M. Bouzinier*, *M. Tsatsulin*, *S. Richards*, *K. Lee*, *C. Almond*. "Mining medical texts for cancer intelligence using iKnow". Cancer Outcomes Conference in 2013.
8. *Bronselaer G.* De Tre, "Concept-Relational Text Clustering", international journal of intelligent systems, 2012, vol. 27, 970–993 (2012).
9. *De Boe M.*, *Bouzinier D.*, *Van Hyfte*. "Extending the PMML Text Model for Text Categorization", PMML workshop @ KDD13, August November 2013, Chicago, Illinois, USA.
10. *Savchenko M.* Auto Semantic Knowledge Network Builder. 2017, https://github.com/ZitRos/edu-semantic-knowledge-network-auto-builder.

Received 21.02.2018

М.М. Савченко, студент (магістр), zitros.lab@gmail.com
О.С. Крячок, к.т.н., доцент, alexandrkriachok@gmail.com

Національний технічний університет України «Київський політехнічний інститут ім. Ігоря Сікорського». м. Київ, пр. Перемоги, 37, корп. 5, 03056.

АВТОМАТИЧНА ПОБУДОВА СЕМАНТИЧНИХ МЕРЕЖ ЗНАНЬ ІЗ НЕСТРУКТУРОВАНИХ ТЕКСТІВ

Вступ. У сучасному інформаційному просторі через величезну кількість неструктурованої текстової інформації існує потреба пошуку, вилучення, формалізації та обробки найбільш суттєвих знань, закладених автором у текст. Такими знаннями можуть бути концепти, представлені в документах, та характерні відносини між ними. Кожний текст будь-якого текстового корпусу несе певний унікальний зміст, характерний лише для даного тексту. Актуальною задачею є розробка алгоритмічної та програмної бази, яка дозволяла б обробляти лише найбільш змістовну частину текстів та вилучати з неї знання, релевантні для даного контексту.

Мета статті. Створення алгоритмічної і програмної бази для побудови семантичних мереж знань з найбільш релевантної інформації відносно контексту документів.

Методи. Запропоновано комплексну методику, алгоритм та його реалізацію для побудови семантичної мережі знань з найбільш значної інформації у заданих текстах. Запропонований комплексний алгоритм поєднує роботу кількох алгоритмів на основі нейронних мереж та статистичного аналізу. Комбінація даних алгоритмів дозволяє розпізнавати концепти в тексті, знаходити між ними зв'язки та визначати, які з концептів мають бути включені до результуючої семантичної мережі шляхом оцінки їх ваги.

Результат. Проведено аналіз великого текстового корпусу, загальною чисельністю близько мільйона слів. На основі зібраної інформації за використання розробленого алгоритму і рекурсивної граматики природної мови побудовано семантичну мережу знань для декількох текстів і окрему поєднану семантичну мережу знань. Проведено порівняння недоліків і переваг розробленого алгоритму відносно кількох існуючих підходів вилучення знань з текстів. Продемонстровано результати.

Висновок. Комплексний метод побудови семантичних мереж поєднує всі переваги описаних в статті методів і не наслідує їх основних недоліків. Комплексним методом можна будувати семантичні мережі (графи знань) з текстів у повністю автоматичному режимі та без втручання експертів. Результати вилучення з текстів основної, найбільш релевантної інформації, представленої у вигляді графу знань можна використовувати в подальшому, наприклад, для розробки систем автоматичного інтелектуального аналізу будь-яких текстових даних.

***Ключові слова**: побудова семантичних мереж, вилучення знань, моделі знань, обробка природної мови*

Н.Н. Савченко, студент (магистр), zitros.lab@gmail.com
А.С. Крячок, к.т.н., доцент, alexandrkriachok@gmail.com

Национальный технический университет Украины «Киевский политехнический институт им. Игоря Сикорского». г. Киев, пр. Победы, 37, корпус 5, 03056.

АВТОМАТИЧЕСКОЕ ПОСТРОЕНИЕ СЕМАНТИЧЕСКОЙ СЕТИ ЗНАНИЙ ИЗ НЕСТРУКТУРИРОВАННЫХ ТЕКСТОВ

Введение. В связи с наличием в современном информационном пространстве огромного количества неструктурированной текстовой информации существует потребность в поиске, изъятии, формализации и обработке наиболее существенных знаний, заложенных авторами в тексты. Такими знаниями могут быть концепты, представленные в документах, и характерные отношения между ними. Каждый текст любого текстового корпуса несёт определённый уникальный смысл, характерный только для данного текста. Актуальная задача — разработка алгоритмической и программной базы, позволяющей обрабатывать только наиболее содержательную часть текстов и извлекать из нее знания, релевантные для данного контекста.

Цель статьи. Создание алгоритмической и программной базы для построения семантических сетей знаний из релевантной относительно контекста документов информации.

Методы. Предложены комплексная методика, алгоритм и его реализация для построения семантической сети знаний из самой значимой информации в заданных текстах. Данный комплексный алгоритм сочетает работу нескольких алгоритмов на основе нейронных сетей и статистического анализа. Комбинация этих алгоритмов позволяет распознавать концепты в тексте, находить между ними связи и определять, какие из концептов должны быть включены в результирующую семантическую сеть с помощью оценки их веса в заданном контексте.

Результат. Проведен анализ большого текстового корпуса, общей численностью около миллиона слов. На основе собранной информации с помощью разработанного алгоритма и рекурсивной грамматики естественного языка построена семантическая сеть знаний для нескольких текстов и отдельная совмещенная семантическая сеть знаний. Проведено сравнение недостатков и преимуществ разработанного алгоритма относительно нескольких существующих подходов извлечения знаний из текстов. Продемонстрированы результаты.

Выводы. Комплексный метод построения семантических сетей сочетает все преимущества описанных методов и не наследует их основных недостатков. Комплексным методом можно строить семантические сети (графы знаний) из текстов в полностью автоматическом режиме без вмешательства экспертов. Результаты извлечения из текстов основной, наиболее релевантной информации, представленной в виде графа знаний, можно использовать в дальнейшем, например, для разработки систем автоматического интеллектуального анализа любых текстовых данных.

***Ключевые слова**: построение семантических сетей, извлечение знаний, модели знаний, обработка естественного языка*