(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
6 May 2010 (06.05.2010)

PCT

(10) International Publication Number
WO 2010/050844 A1

(51) International Patent Classification:
G06F 17/27 (2006.01)

(21) International Application Number:
PCT/RU2009/000111

(22) International Filing Date:
6 March 2009 (06.03.2009)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
2008142648 29 October 2008 (29.10.2008) RU

(71) Applicant *(for all designated States except US)*: **ZAKRYTOE AKTSIONERNOE OBSCHESTVO "AVICOMP SERVICES"** [RU/RU]; prosp. Vernadskogo, 84-2, Moscow, 119606 (RU).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **KHOROSHEVSKY, Vladimir Fyodorovich** [RU/RU]; prosp. Vernadskogo, 15-309, Moscow, 119311 (RU). **KLINTSOV, Victor Petrovich** [RU/RU]; ul. Gorkogo, 12, pos. Kratovo, Ramensky raion, Moskovskaya obl., 140133 (RU).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report (Art. 21(3))*

(54) **Title**: METHOD OF COMPUTERIZED SEMANTIC INDEXING OF NATURAL LANGUAGE TEXT, METHOD OF COMPUTERIZED SEMANTIC INDEXING OF COLLECTION OF NATURAL LANGUAGE TEXTS, AND MACHINE-READABLE MEDIA

(57) **Abstract**: The present invention relates to the information technologies field, namely, to methods of computerized semantic indexing of natural language texts. The use of the present invention permits for extending the set of methods for indexing the natural languages texts by means of employing techniques of the computerized linguistic analysis thereof and further usage of obtained results for building indices, which ensures the semantic navigation through documents and document collections, as well as the highly-precise and quick search of facts and documents relevant to the user's information needs, particularly, in reference to the high-inflectional language texts. The method of computerized semantic indexing of natural language text comprises steps of: segmenting the text in the electronic form into tokens; identifying stable phrases; forming sentences; by addressing the linguistic and heuristic rules formed in the database in the predetermined linguistic environment, identifying the semantically meaningful objects (named entities) and the semantically meaningful relations therebetween (named relations); for every named relations, forming the set of triples, where single first type triple corresponding to the relation established by the named relation between two named entities, each of the set of the second type triples corresponding to a value of particular attribute of one of those entities, and each of the set of the third type triples corresponding to a value of particular attribute of the named relation itself; at the set of the formed triples, indexing all named entities related by the named relations separately, all pairs of the kind "named entity - named relation", and all triples of the kind "named entity — named relation - named entity", while taking into account the attributes of respective named entities and/or named relations; and storing in the database the formed triples and the obtained indices together with the reference to the initial text from which those triples have been formed.

WO 2010/050844 A1

METHOD OF COMPUTERIZED SEMANTIC INDEXING OF NATURAL LANGUAGE TEXT, METHOD OF COMPUTERIZED SEMANTIC INDEXING OF COLLECTION OF NATURAL LANGUAGE TEXTS, AND MACHINE-READABLE MEDIA

Field of the Invention

The present invention relates to the information technologies field, namely, to methods of computerized semantic indexing of natural language texts, as well as to machine-readable media comprising respective programs, and could be used for ordering and accumulating information in specified knowledge areas for the purpose of semantic navigation through the documents and document collections, as well as for the highly-precise and quick search of facts and documents relevant to the user's information needs.

Background of the Invention

At present, various methods of the computerized indexing of natural language texts are known.

So, the EAPO Patent No. 002016 (published on 2001.01.22) describes a method, where unique information blocks are detected in text document fragments and used for subsequent processing and searching. The RU Patent No. 2268488 (published on 2006.01.20) granted on the basis of the PCT Application published as WO 01/06414 discloses the method wherein words, stable phrases, idioms, sentences, and even ideas are coded for the subsequent processing at the numerical level. The RU Patent No. 2273879 (published on 2006.04.10) adduces a method wherein morphological and syntactic text analysis with the subsequent indexing of the detected units. In the method of the US Patent No. 6,871,174 (published on 2005.03.22), a text similarity is determined by text fragments. The

disadvantage of all those methods consists in that they do not take into account the semantic ambiguity of the natural language words and expressions.

The US Patent No. 6,189,002 (published on 2001.02.13) discloses a method, wherein a text is divided into paragraphs and words that are converted into vectors of the ordered element. Each vector element corresponds to the paragraph determined by applying the predetermined function to a number of occurrences of the word corresponding to that element, in this paragraph. The text vector is considered as the semantic profile of the document. However, taking into account the variety of paragraphs, this method requires an enormous massive of the stored data and does not distinguish semantic ambiguity of the words and expressions.

The semantic ambiguity consideration is carried out in many known methods. For example, the RU Patent No. 2242048 (published on 2005.03.22), US Patents Nos. 6,871,199 (published on 2005.03.22), 7,024,407 (published on 2006.04.04), and 7,383,169 (published on 2008.06.03), US Patent Application Publications Nos. 2007/0005343 and 2007/0005344 (published both on 2007.01.04), 2008/0097951 (published on 2008.04.24), JP Laid-open Applications Nos. 05-128149 (published on 1993.05.25), 06-195374 (published on 1994.07.15), 10-171806 (published on 1998.06.26), and 2005-182438 (published on 2005.07.07), EP Patent Application N0. 0853286 (published on 1998.07.15) methods are described, wherein the ambiguity of the words and/or expressions being found in texts is eliminated in one or another manner. However, all those methods have only restricted application and do not affect a valuable semantic text indexing.

Closest to the claimed group of inventions is a method of computerized semantic indexing of natural language text or text collection, which method is disclosed in the US Patent Application No. 2007/0073533 (published on 2007.03.29). In that method, functional structures are determined for every text portion, and triples characterizing predicate members are determined in every

functional structure based on the linearization transfer rules. Then, such features as named entity, co-reference, lexical entry, structural-semantic relationship, and attribution and meronymic information are extracted from every text portion. Next, for every text portion, on the basis of the determined constituent structure, canonical triple representations characterizing the predicative members are determined together with the derived feature representations, and a structural index is determined on the basis of the canonical representation of the text portion. This method ensures good results, but yet is rather limited due to the fact that fragments of the predicative-argument structure derived during the syntactic analysis are linearized as the triples. Moreover, this method is directed to the search tasks only, rather than the tasks of navigation through a document collection.

Summary of the Invention

The object of the present invention consists in extending the set of methods for indexing the natural languages texts by means of employing techniques of the computerized linguistic analysis thereof and further usage of obtained results for building indices, which ensures the semantic navigation through documents and document collections, as well as the highly-precise and quick search of facts and documents relevant to the user's information needs, particularly, in reference to the high-inflectional language texts.

An achievement of this object and obtaining of said technical result is provided using a method of computerized semantic indexing of natural language text and a method of computerized semantic indexing of natural language text collection in accordance with the features of the enclosed independent claims 1 and 7, respectively. The variations of the both methods are disclosed in the respective dependent claims.

Brief Description of the Drawings

The invention is explained by describing a particular embodiment thereof and accompanying drawings, in which:

Fig. 1 depicts the schematic block-diagram explaining the claimed methods;

Fig. 2 depicts the fragment of the application domain specification;

Fig. 3 shows the rule schema for extracting the named entities of the type of "Person";

Fig. 4 shows the rule schema for extracting the semantically meaningful relationships of the type of "work";

Fig. 5 depicts the fragment of the graphical form for representing the results of the text processing;

Fig. 6 shows the general diagram for storing the results of processing one text;

Fig. 7 shows the diagram of the left-hand side of the rule for combining the named entities of the type of "Person".

Detailed Description of the Invention

The proposed methods allow for performing effectively the conceptual indexing of the natural language texts both for the subsequent semantic navigation through the documents and document collections and for search purposes.

The method of computerized semantic indexing of natural language text according to the first aspect of the present invention and the method of computerized semantic indexing of collection of natural language text according to the second aspect of the present invention could be implemented in practically either computing environment, e.g., in a personal computer connected to external databases. The steps of performing those methods are illustrated in Fig. 1.

All further explanations are adduced in reference to the Russian language which is one of most high-inflectional languages, although the claimed methods are applicable to the semantic text indexing in any natural languages.

First of all, for the subsequent computerized treatment, the text to be indexing needs to be presented in an electronic form. This step in Fig. 1 is contingently marked with the reference 1 and could be performed by any known method, for example, by scanning the text with recognizing thereof subsequently using a well-known means of the type of ABBYY FineReader. If the text comes to the indexing from the electronic network, for example, from the Internet, then the step of representing thereof in the electronic form is carried out in advance, prior to disposing that text in the network.

The text converted into the electronic form comes to the processing, during which this text is first segmented into elementary units named tokens. A token could be any text object from the following set: words consisting each of the series of letters and, possibly, hyphens; a series of spaces; punctuation marks; numbers. Sometimes, such character series as A300, i150b, etc. are also pertained hereinto. Tokens' separation is always carried out in accordance with rather simple rules, for example, as in the mentioned US Patent Application No. 2007/0073533. In Fig. 1, this step is contingently marked with the reference number 2. Hereinafter, tokens are considered as the first level elementary units.

Moreover, for every token being a word, respective second level elementary units named hereinafter morphs are formed based on the morphological analysis. In so doing, for every word, its normalized word form is identified. For example, for the word «иду» the normalized word form will be «идти», for the word «красивого» the normalized word form will be «красивый», and for the word «стена» the normalized word form will be «стена». Moreover, for every word form, a part of speech to which this word relates and its morphological characteristics are indicated. Of course, for various parts of speech those characteristics differ. For example, for nouns and adjectives these are a gender (masculine, feminine, neuter), number (singular, plural), case; for verbs these are aspect (perfective, imperfective), person, number (singular, plural); etc. Thus, for the given word, its morph is the normalized word form + morphological characteris-

tics including the part of speech. One and the same word can have several word forms. For example, the word «стекло» has two morphs: one for the neuter noun ["glass"] and one for the verb in the past tense ["flowed down"].

Those skilled in the art would appreciate that operations of this and subsequent steps are carried out with storing the intermediate results, for example, in the random access memory (RAM).

The next step contingently marked with the reference number 3 in Fig. 1 consists in that stable phrases are identified in a set of derived elementary units of two first levels, tokens and morphs. This is performed by converting the elementary units, i.e., tokens and morphs, into series that are compared with the series of normalized words and their characteristics in the dictionaries stored in advance in the databases, where words are adduced with specification of grammatical associations therebetween. Once coinciding the succeeding series being compared with the corresponding dictionary series, that succeeding series being compared is recognized as the stable phrase and stored in such a kind in the database as the third level elementary unit.

At the next step indicated by the reference number 4 in Fig. 1, sentences corresponding to the portions of the text being indexed are formed. Usually, these are real sentences ending with dot, but in some cases it is suitable for interpreting as a sentence some parts of usual sentences, say, isolated element in enumeration. Therefore, this step can issue sentences not always coincided with the sentences of the text being indexed in common sense.

The above series of steps is specified by that identifying the lookup prior to forming the sentences allows, in some instances, to eliminate certain ambiguousness before analyzing the text in details. Thus, for example, fixation of the lookup «МГУ им. М.В. Ломоносова» ["Moscow State University named for M.V. Lomonosov"] allows to eliminate the false end of the sentence after the word «им» which is, in general case, the pronoun ["them"], but here is the shortening of the word «имени» ["named for"].

After the step 4, the multistage semantic-syntactic analysis is carried out. In Fig. 4, this analysis is contingently divided into steps marked with the reference numbers 5 to 11. Said multistage semantic-syntactic analysis is carried out by addressing the linguistic and heuristic rules formed in the database in the predetermined linguistic environment. Such an environment could be, for example, the linguistic environment mentioned in the above RU Patent No. 2242048, or the environment disclosed in said US Patent Application No. 2007/0073533, or any other linguistic environment defining respective rules that allows to eliminate syntactic and semantic ambiguousness of the real text words and expressions. The linguistic and heuristic rules in the chosen environments are hereinafter referred to as rules. During said multistage semantic-syntactic analysis, the semantically meaningful objects hereinafter referred to as named entities (the reference number 5 in Fig. 1), and the attributes thereof (the reference numbers 7 and 9 in Fig. 1) are identified.

The identification of the named entities that are considered as the fourth level elementary units is carried out in the sentence in a set of elementary units of the first, second, and/or third levels. In this case, the morphological attributes are formed for every named entity using said rules from the morphological attributes of those elementary units of the second and/or third levels (i.e., morphs and/or stable phrases) which constitute this named entity. Moreover, the semantic attributes are formed for every named entity using said rules from the semantic attributes of the elementary units of the second and/or third levels which constitute this named entity. The step of forming said attributes is contingently marked with the reference 7. And at the step marked with the reference number 9 in Fig. 1, for every named entity is assigned a respective type from the application ontology according to the topics of the application domain, to which the text being indexed relates. By the application ontology is meant, in this case, the specification of the particular application domain, which is stored in the respective database.

For every named entity, i.e., for the fourth level elementary unit with type assigned thereto, the corresponding anaphoric reference considered as the fifth level elementary unit (if any) is determined. For example, in the sentence «Когда Путина назначили премьер-министром, он сформировал правительство» ["When Putin was appointed to the post of prime-minister, he has formed the government"], the anaphoric reference to the word «Путин» [«Путина»] will be the pronoun «он» ["he"], while the word «Путин» will be the antecedent for that anaphor. This step for determining the anaphoric reference is contingently marked with the reference number 11 in Fig. 1.

After that, every identified named entity is stored in the respective memory together with the type assigned thereto and morphologic and semantic attributes determined thereto. The anaphoric reference is stored together with the type and attributes of the named entity which is the antecedent of that anaphoric reference, as well as with the indication of the co-reference between that named entity and the anaphoric reference thereof.

After performing the steps marked with the reference numbers 5-7-9-11 in Fig. 1, the semantically meaningful relations between the named entities hereinafter referred to as named relations are determined based on the elementary units of the first, second, third, fourth and/or fifth levels using said rules (the step 6). The named relations can relate the named entities within both one sentence and the entire text being indexed.

At the step marked with the reference number 8 in Fig. 1, the morphological attributes are determined for every named relation using said rules from the second level elementary units (i.e., morphs) constituting this relation, as well as the semantic attributes from the elementary units of the first, second, third and/or fourth levels, constituting this relation.

At the step marked with the reference number 10 in Fig. 1, the respective type is assigned to every named relation from application ontology stored in the database according to the topics of the application domain, to which the text be-

ing indexed relates. After that, every named relation is stored in the respective memory together with the type assigned thereto and morphologic and semantic attributes determined thereto.

At the step marked with the reference number 12 in Fig. 1, the stored named entities and named relations are used for forming the triples. In so doing, a set of the triples of three types is formed within the text being indexed for every of the identified named relations relating the certain named entities. The single first type triple corresponds to the relation established by the named relation between two named entities. Each of the set of the second type triples corresponds to a value of particular attribute of one of those entities, and each of the set of the third type triples corresponds to a value of particular attribute of the named relation itself. If two named entities are labeled by O_i and O_j, and the named relation relating thereof is labeled as R_{ij}, then the first type triple could be represented (depicted) as $O_i \rightarrow R_{ij} \rightarrow O_j$. Each of the set of the second type triples could be represented as $O_i \rightarrow A_{im} \rightarrow V_{im}$ or $O_j \rightarrow A_{jn} \rightarrow V_{jn}$, where A_{im} and A_{jn} are respective attributes, V_{im} and V_{jn} are, respectively, values of those attributes. Similarly, each of the set of the third type triples could be represented as $R_{ij} \rightarrow A_{ijk} \rightarrow V_{ijp}$, where A_{ijk} is a respective attribute, and V_{ijp} is a value of that attribute. In these notations, the indices i, j, k, m, n, and p are integers.

Then, at the step marked with the reference number 13 in Fig. 1, the text indexing is carried out. For this purpose, all named entities related by the named relations separately, all pairs of the kind "named entity – named relation", and all triples of the kind "named entity – named relation – named entity" are indexed taking into account the attributes of respective named entities and/or named relations. The triples formed at the step 12 and indices obtained at the step 13, together with the reference to the initial text from which those triples have been formed, are stored in the database (the step 15 in Fig. 1; the step 14 is omitted in this case). Prior to this, a convolution is performed (not shown) for the objects related by co-reference relations into a single object whose set of the

attributes are the combination of the attributes of all object interrelated by the co-reference relations. This is done in order for reducing the memory volume in the database required for storing such objects, as well as for integrating under one object the information obtained for the entire text.

The method of computerized semantic indexing of collection of natural language texts according to the second aspect of the present invention is carried out exactly as already discussed method of computerized semantic indexing of natural language texts according to the first aspect of the present invention, but in this case, after the step 13 of indexing and prior to the step 15 of storing in the database, one more step is performed. At this step marked with the reference number 14 in Fig. 1 and performed substantially simultaneously with the step 15, the following is carried out when storing in the database the formed triples and obtained semantic indices of the succeeding text. The newly derived named entities and named relations are compared with the named entities and named relations already existed in the database using the linguistic and heuristic rules in the predetermined linguistic environment that are formed in the database. In the case of identifying similar named entities and/or named relations, the duplicated information is not stored in the database, and respective named entities and/or named relations are supplemented with references to the succeeding texts where they are present and references to the text fragments within each of succeeding texts from which they are derived. Through this, the step of indexing the text collection is occurred similarly to the indexing the first text of this collection (or the first text indexed by this method), which permits to simplify significantly the entire indexing procedure, reduce the required memory volume, and integrate the information obtained from different texts within a single object.

It should be apparent for those skilled in the art that storage devices mentioned at individual steps could be practically both different devices and a single storage device of sufficient volume. Similarly, individual databases mentioned at the respective steps could be not only physically separate databases, but also a

single database. Furthermore, said storage devices (memories) could be made on the same single database, or combined with one of said databases. Those skilled in the art will also appreciate that the methods claimed in the present invention are carried out in the respective computing environment under the control of appropriate programs which are recorded on machine-readable media intended for a direct utilization in computer operation. Therefore, the aspects of the present invention are also the machine-readable media with such programs.

Example

In order for illustrating the embodiment of the claimed method of computerized semantic indexing of natural language texts, consider the following example. Let there is a set of Russian texts formed from the newslines of the electronic mass media. Thus, the step of converting the texts into the electronic form marked in Fig. 1 with the reference number 1 could be considered to be already performed.

A representative example of such text is the following message:

«Центральный федеральный округ / Публикации за 26.06.08

Ющенко поручил Тимошенко выведать у Путина цену на газ

26.06.08 14:00

Киев, Июнь 26 (Новый Регион, Михаил Рябов) – Президент Украины Виктор Ющенко заявил, что ещё вчера утвердил директивы для переговоров премьер-министра Юлии Тимошенко с российским коллегой Владимиром Путиным в Москве 28 июня.

По словам Ющенко, Тимошенко во время переговоров с Путиным должна определить «чёткую формулу цены на газ», причём, «чтобы это была цена не политическая, это была цена экономическая».

«Это значит, что обеим сторонам понятны зависимости, на базе которых формируется цена на газ в очередном году», – пояснил Ющенко.

Он дал понять, что в случае повышения цены на газ Украина будет добиваться пересмотра стоимости транзита российского газа в Европу.

«Тут должна быть коррекция и увязка политики цены на газ и политики цены на тариф», – заявил президент Украины.

По словам Ющенко, Тимошенко должна добиться, чтобы формула цены на газ для Украины в 2009 году стала известна не позднее 15 сентября 2008 года, чтобы украинская сторона смогла учесть новые расценки в госбюджете на следующий год.

Как сообщал «Новый Регион», ранее глава МИД России Сергей Лавров заявлял, что цена на газ для Украины в 2009 году может вырасти вдвое.».

["Central Federal District / Publication from 06.26.2008

Uschenko has entrusted Timoshenko to ferret from Putin the price for gas.

06.26.2008, 2:00 PM

Kiev, June 26 (New Region, Michail Ryabov) – The Ukrainian President Victor Uschenko has intimated that yet yesterday he has approved the guide lines for negotiations of the Prime Minister Yulia Timoshenko with the Russian colleague Vladimir Putin in Moscow on June 28.

According to Uschenko, Timoshenko, during the negotiations with Putin, must define "clean formula of the price for gas", herewith "this price will be economical price rather than political price".

"This means that both sides understand the dependencies on which basis the price for gas is formed in the next year", has explained Uschenko.

He has given to understand that, in the case of increasing the price for gas, Ukraine will hold out for revising the cost for transiting Russian gas to Europe.

"Here must be correcting and matching the gas prices policy and tariff price policy", has stated the Ukrainian President.

According to Uschenko, Timoshenko must enforce that the formula of the price for gar for Ukraine in 2009 becomes known not later than on September 15, 2008, in order for the Ukrainian party could take into account the new rates in the state budget for the next year.

As the "New Region" informed, the head of the Russian Foreign Office Sergey Lavrov told earlier the price for gas for Ukraine in the 2009 could be doubled.".]

In accordance with the claimed method of computerized semantic indexing of natural language texts, the preliminary created application domain specification is used, within which the text collection processing and semantic index constructing will be carried out. A fragment of such specification is depicted in Fig. 2. Such specifications are developed by human experts, who record, based on their experience and knowledge, a list of object types and a list of typical relations therebetween essential for this application domain.

In the example cited, the main types of objects are "Person", "Organization", "Location", and some other. Typical relations therebetween fall into two classes: common, peculiar to any domains, for example, the relation "BE_EXAMPLIFIED", which indicate the object hierarchy of the type of "descendant-ancestor", and special, specific for the chosen application domain, for instance, in the cited example, this is the typical relations "work ", "own", "located_in", etc.

Moreover, the human experts build in advance also a set of rules, each rule containing, in the left-hand side, a template for searching examples of objects and/or examples of relations therebetween, and in the right-hand side, operators for fixing in the text the examples of objects and/or examples of relations therebetween determined in accordance with the template. Using those rules prepared by the human experts, the specific data corresponding to the domain specification are derived in the texts being processed.

Besides the domain specification and the rules in accordance with the above methods, common and special lexicons are used.

In accordance with the claimed method of computerized semantic indexing of natural language texts, at first, the step of segmenting the text into elementary units, tokens, is performed with the morphological analysis of the token-words (reference number 2 in Fig. 1). As a result of this step, the initial text is transformed into a set of tokens and morphs that are represented in the Table 1 and Table 2, respectively.

Further, after performing the text segmentation into the tokens and the morphological analysis of the token-words, the step is carried out for deriving the stable phrases (lookups) using the common and special lexicons (reference number 3 in Fig. 1). As a result of this step, the initial text is supplemented, besides the first and second level elementary units, with a set of the third level elementary units, lookups. The fragment of this set for the above example is represented in the Table 3.

After performing the above steps, the text being processed is fragmented into sentences (reference number 4 in Fig. 1). As a result of this step, the pluralities formed at the above steps are supplemented with a set of sentences, represented in the Table 4.

Thus, after performing all the above discussed steps, the text being processed will be segmented into the sentences, each of which is marked with a plurality of annotations of the first, second and third level elementary units.

Hereupon, in accordance with the claimed method of computerized semantic indexing of natural language texts, the step of deriving the named entities (fourth level elementary units) is carried out at the set of the elementary units of the first, second and/or third levels using said rules. Thus, for example, in the sentence «Как сообщал «Новый Регион», ранее глава МИД России Сергей Лавров заявлял, что цена на газ для Украины в 2009 году может вырасти вдвое.» [As the "New Region" informed, the head of the MFA of Russia Sergey

Lavrov told earlier the price for gas for Ukraine in the 2009 could be doubled."] of the text being considered, the named entities «Новый Регион» ["New Region"], «Сергей Лавров» ["Sergey Lavrov"], «Украина» ["Ukraine"], and some others are derived using the rule whose diagram is represented in Fig. 3. As a result of performing the steps marked with the reference numbers 5-7-9 in Fig. 1, the fourth level elementary units (named entities together with the attributes thereof) are derived. The fragment of plurality of such units for the example under consideration is represented in the Table 5.

After that, within the entire text being processed, during the step marked in Fig. 1 with the reference number 11, the pronouns are determined that could be anaphoric references to the corresponding named entities, and for those pronouns that are really such ones, the co-reference between the respective named entity and the anaphoric reference thereof (the fifth level elementary unit) is fixed. For the example under consideration, the obtained set of the anaphoric references is represented in the Table 6.

After performing the above steps, at the set of the derived first, second, third, fourth and fifth level elementary units, the semantically meaningful relations between the named entities are determined using the rules. Thus, for example, in the sentence «Как сообщал «Новый Регион», ранее глава МИД России Сергей Лавров заявлял, что цена на газ для Украины в 2009 году может вырасти вдвое.» [As the "New Region" informed, the head of the MFA of Russia Sergey Lavrov told earlier the price for gas for Ukraine in the 2009 could be doubled."] of the text being considered, the named relation «работать» ["work"] is derived using the rule whose diagram is represented in Fig. 4. As a result of performing the steps marked with the reference numbers 6-8-10 in Fig. 1, the set of the named relations between the named entities is derived, the fragment of which for the example under consideration is represented in the Table 7.

Thus, after performing all the above discussed steps, the initial text will be marked with the set of annotations corresponding to the named entities with the

attributes thereof and the named relations with the attributes thereof between the named entities. For the example under consideration, the graphical representation of the text processing results is shown in Fig. 5.

The next step marked with the reference number 12 in Fig. 1 is a technical and carried out for performing the triples corresponding to the stored named entities and named relations. The fragment of the set of such triples for the example under consideration is represented in the Table 8. In fact, the formed set of triples contains the initial data for the semantic indexing of the text processed at the previous steps.

At the step marked with the reference number 13 in Fig. 1, the semantic index is built as follows: first, from the set of the triples obtained at the previous step, the triple subsets are formed, each oh which subsets corresponds to one named entity with the attributes thereof, and every obtained triple subset is used as an entry for one of conventional indexers, for example, the well-known, freeware indexer Lucene, the indexer of the Yandex search machine, the Google indexer, or any other indexer, from which output an index unique for the given triple subset is obtained. The similar operation series is performed for all subsets of triples corresponding to the pairs of the kind "named entity – named relation" and to the triples of the kind of "named entity – named relation – named entity" taking into account the attributes of the respective named entities and/or named relations, thereby obtaining a set of the corresponding unique indices which constitute, in the aggregate, the semantic index of the text. The fragment of the semantic index for the example under consideration is represented in the Tables 9 to 11.

At the step marked with the reference number 15 in Fig. 1, the triples formed at the step 12 and the indices obtained at the step 13 together with the reference to the initial text from which those triples have been formed are stored in the database, and the step 14, in the case of processing one text, is omitted.

The general diagram of storing all results obtained at the previous steps is represented in Fig. 6.

Referring to Fig. 6, as the first step (51), a set of continuous chains of the triples for the relation "The_same" are formed. At the next step (52), the check is performed whether the set of continuous chains of the triples obtained at the previous step is empty. If that set is not empty, then, sequentially, at the next steps (53-56), the set of objects for the next chain is formed (53), this set is convolved into the single object (54) having the combined set of the attributes (without repetitions), the obtained single object is stored together with the attributes thereof (55), and the set of the processed objects of the succeeding chain is removed (56). But if at the step 52 the set of the triple chains turns out to be empty (initially or as a result of performing the steps 53 to 55), then, at the step 57, the overall set of triples obtained at all previous steps is formed; at the step 58, the formed overall set of the triples is supplemented with the semantic indices and references to the initial text; after which, at the step 59, the supplemented set is stored in the database.

In accordance with the claimed method of computerized semantic indexing of collection of natural language texts, the processing of every subsequent text including its semantic index constructing is carried out by performing just the same steps as for the single text. However, in this case, after the step 13 of indexing and prior to the step 15 of storing in the database, one more step marked with the reference number 14 in Fig. 1 is carried out, the step of combining the results of processing the succeeding text with the results of processing the previous texts stored already in the database, which step is carried out as follows.

The named objects and named relations newly derived in the succeeding text being indexed are compared with the named objects and named relations already existed in the database by checking the coincidence of the semantic indices thereof, and, in the case of the positive result of such comparing, the respec-

tive objects and relations are excluded from the following processing, herewith storing in the object and/or relation already existed in the database the reference to that text and that fragment of that text, where the object and/or relation excluded from the following processing are identified. In the case of the negative result of comparing the semantic indices, the similarity between the new objects and/or relations and those object and/or relations that are already exist in the database is identified using the linguistic and heuristic rules formed in advance in the database, and in the case of the positive result, the object and/or relation descriptions already exist in the database are widened with the new data, after which the existing semantic indices are reconstituted and the new semantic indices are added as the secondary ones to the already existing indices, and, moreover, in the object and/or relation already existing in the database, the reference is stored to that text and that fragment of this text, where the new objects and/or relations are identified, and then the respective objects and relations are excluded from the following processing. Otherwise, the newly identified named objects and named relations together with the semantic indices thereof are added to the database.

Thus, for example, if, as the next one relative to the just considered example, the following text was processed: «Киев, Июль 21 (Новый Регион, Анна Сергеева) – Сегодня в Киеве состоялись двусторонние переговоры руководства Украины и Германии. Встреча Виктора Ющенко и Ангелы Меркель в формате с глазу на глаз длилась 15 минут, а затем начались переговоры в расширенном формате, которые затянулись на полтора часа... Виктор Андреевич Ющенко выглядел довольно вяло на фоне бойкой и активной Меркель, которая быстро и много говорила...» [“Kiev, July 21 (New Region, Anna Sergeyeva) – Today in Kiev, the bilateral negotiations have been happened between the Ukrainian and German leaders. The meeting of Victor Uschenko and Angela Merkel in vis-a-vis format lasted for 15 minutes, and then the negotiations in the widened format have begun, which lasted for one hour

and half... Victor Andreyevich Uschenko seemed to be rather slow against lively and active Merkel, who spoke quick and much..."], then, after performing the steps 1 to 13, there will be identified, for example, such objects and relations as «Новый Регион» ["New Region"], «Украина» ["Ukraine"], «Виктор Ющенко» ["Victor Uschenko"], «Ангела Меркель» ["Angela Merkel"], «Виктор Андреевич Ющенко» ["Victor Andreyevish Uschenko"], «встречи_переговоры» ["meetings_negotiations"], etc., as well as the semantic index of this text will be formed, which fragment is represented in the Tables 12 to 14.

Further, in accordance with the claimed method of computerized semantic indexing of collection of natural language texts, at the step 14, the object will be identified, particularly, the object «Украина» ["Ukraine"], which semantic index fully coincides with the semantic index of the object «Украина» ["Ukraine"] already existed in the database, and, moreover, the similarity will be identified (by applying the rule which diagram is shown in Fig. 7) between the object «Виктор Андреевич Ющенко» ["Victor Andreyevich Uschenko"] and the object «Виктор Ющенко» ["Victor Uschenko"] already existed in the database, whereafter the existed description of the object «Виктор Ющенко» ["Victor Uschenko"] in the database will be widened by means of the new information and additional semantic index, which is shown in the Tables 15 and 16.

Thus, the present invention provides for extending the set of methods for indexing the natural languages texts by means of employing techniques of the computerized linguistic analysis thereof and subsequent use of obtained results for building indices, the main difference of which methods from the known method of indexing consists in indexing semantically meaningful concepts and relations therebetween rather than the key words and lookups, which provides for the semantic navigation through the documents and documents collections, as well as the highly-precise and quick search of facts and documents, especially in reference to high-inflectional language texts.

Table 1. Results of tokenizing the example text

Element type	Start position	End position	Attributes
Token	0	11	{type=word, coding=cyr, length=11, orpho=upperInitial, string=Центральный [Central]}
Token	12	23	{type=word, coding=cyr, length=11, orpho=lowercase, string=федеральный [federal]}
Token	24	29	{type=word, coding=cyr, length=5, orpho=lowercase, string=округ [district]}
.....			
Token	57	63	{type=word, coding=cyr, length=6, orpho=upperInitial, string=Ющенко [Uschenko]}
Token	64	71	{type=word, coding=cyr, length=7, orpho=lowercase, string=поручил [has charged}
Token	72	81	{type=word, coding=cyr, length=9, orpho=upperInitial, string=Тимошенко [Timoshenko]}
Token	82	90	{type=word, coding=cyr, length=8, orpho=lowercase, string=выведать [find]}
Token	91	92	{type=word, coding=cyr, length=1, orpho=lowercase, string=у [out]}
Token	93	99	{type=word, coding=cyr, length=6, orpho=upperInitial, string=Путина [Putin]}
Token	100	104	{type=word, coding=cyr, length=4, orpho=lowercase, string=цену [price]}
Token	105	107	{type=word, coding=cyr, length=2, orpho=lowercase, string=на [for]}
Token	108	111	{type=word, coding=cyr, length=3, orpho=lowercase, string=газ [gas]}
.....			
Token	1176	1181	{type=word, coding=cyr, length=5, orpho=lowercase, string=глава [head]}
Token	1182	1185	{type=word, coding=cyr, length=3, orpho=allCaps, string=МИД {MFA}
Token	1186	1192	{type=word, coding=cyr, length=6, orpho=upperInitial, string=России [Russia]}
Token	1193	1199	{type=word, coding=cyr, length=6, orpho=upperInitial, string=Сергей [Sergey}

Token	1200	1206	{type=word, coding=cyr, length=6, orpho=upperInitial, string=Лавров [Lavrov]}
Token	1207	1214	{type=word, coding=cyr, length=7, orpho=lowercase, string=заявлял [announced]}
Token	1214	1215	{type=punctuation, length=1, string=,}
Token	1216	1219	{type=word, coding=cyr, length=3, orpho=lowercase, string=что [that]
Token	1220	1224	{type=word, coding=cyr, length=4, orpho=lowercase, string=цена [price]}
Token	1225	1227	{type=word, coding=cyr, length=2, orpho=lowercase, string=на [for]}
Token	1228	1231	{type=word, coding=cyr, length=3, orpho=lowercase, string=газ [gas]}
Token	1232	1235	{type=word, coding=cyr, length=3, orpho=lowercase, string=для [for]}
Token	1236	1243	{type=word, coding=cyr, length=7, orpho=upperInitial, string=Украины [Ukraine]}
Token	1244	1245	{type=word, coding=cyr, length=1, orpho=lowercase, string=в [in]}
Token	1246	1250	{type=number, length=4, string=2009}
Token	1251	1255	{type=word, coding=cyr, length=4, orpho=lowercase, string=году [year]}
Token	1256	1261	{type=word, coding=cyr, length=5, orpho=lowercase, string=может [maybe]}
Token	1262	1269	{type=word, coding=cyr, length=7, orpho=lowercase, string=вырасти [increase]}
Token	1270	1275	{type=word, coding=cyr, length=5, orpho=lowercase, string=вдвое [doubled]}
Token	1275	1276	{type=punctuation, length=1, string=.}

Table 2. Results of the morphological analysis of the example text

Element type	Start position	End position	Attributes
Morph	0	11	{CAS=acc, GEND=m, NMB=sg, POS=A, base=центральный [central]}
Morph	0	11	{CAS=nom, GEND=m, NMB=sg, POS=A, base=центральный [central]}
Morph	12	23	{CAS=acc, GEND=m, NMB=sg, POS=A, base=федеральный [federal]}
Morph	12	23	{CAS=nom, GEND=m, NMB=sg, POS=A, base=федеральный [federal]}
Morph	24	29	{CAS=gen, POS=PREP, base=округ [district]}
.....			
Morph	24	29	{CAS=nom, GEND=m, NMB=sg, POS=N, base=округ [district]}
Morph	24	29	{CAS=gen, GEND=f, NMB=pl, POS=N, base=округа [district]}
.....			
Morph	1176	1181	{CAS=nom, GEND=m, NMB=sg, POS=N, base=глава [head]}
Morph	1176	1181	{CAS=nom, GEND=f, NMB=sg, POS=N, base=глава [head]}
Morph	1182	1185	{CAS=dat, GEND=n, INVAR=invar, NMB=sg, POS=N, base=МИД [MFA], freqOrthm=u*}
.....			
Morph	1182	1185	{CAS=prp, GEND=n, INVAR=invar, NMB=pl, POS=N, base= МИД [MFA], freqOrthm=u*}
Morph	1186	1192	{CAS=dat, GEND=f, NMB=sg, POS=N, base=Россия [Russia], freqOrthm=u*}
Morph	1186	1192	{CAS=gen, GEND=f, NMB=sg, POS=N, base=Россия [Russia], freqOrthm=u*}
Morph	1186	1192	{CAS=prp, GEND=f, NMB=sg, POS=N, base=Россия [Russia], freqOrthm=u*}
Morph	1193	1199	{CAS=nom, GEND=m, NMB=sg, PNT=fnam, POS=N, base=Сергей [Sergey], freqOrthm=u*}
Morph	1200	1206	{GEND=m, NMB=sg, POS=A, base=лавровый [laurel]}
Morph	1200	1206	{CAS=gen, GEND=m, NMB=pl, POS=N,

			base=лавр [laurel]}
Morph	1200	1206	{CAS=nom, GEND=m, NMB=sg, PNT=fam, POS=N, base=Лавров [Lavrov], freqOrthm=u*}
Morph	1207	1214	{GEND=m, MD=ind, NMB=sg, POS=V, REPR=fin, TNS=past, TRANS=vt, base=заявлять [announce]}
Morph	1216	1219	{POS=CONJ, base=что}
Morph	1216	1219	{CAS=acc, GEND=n, NMB=sg, POS=N, PRN=prn, base=что [that]}
Morph	1220	1224	{CAS=nom, GEND=f, NMB=sg, POS=N, base=цена [price]}
Morph	1225	1227	{CAS=prp, POS=PREP, base=на [for]}
Morph	1225	1227	{CAS=acc, POS=PREP, base=на [for]}
Morph	1228	1231	{CAS=nom, GEND=m, NMB=sg, POS=N, base=газ [gas]}
Morph	1228	1231	{CAS=acc, GEND=m, NMB=sg, POS=N, base=газ [gas]}
Morph	1232	1235	{CAS=gen, POS=PREP, base=для [for]}
Morph	1232	1235	{POS=V, REPR=gern, TNS=pres, TRANS=vt, base=длить [last]}
Morph	1236	1243	{CAS=gen, GEND=f, NMB=sg, POS=N, base=Украина [Ukraine], freqOrthm=u*}
Morph	1236	1243	{CAS=nom, GEND=f, NMB=pl, POS=N, base=Украина [Ukraine], freqOrthm=u*}
Morph	1236	1243	{CAS=acc, GEND=f, NMB=pl, POS=N, base=Украина [Ukraine], freqOrthm=u*}

.....

Morph	1256	1261	{MD=ind, NMB=sg, POS=V, REPR=fin, TNS=pres, base=мочь [maybe]}
Morph	1262	1269	{MD=imp, NMB=sg, POS=V, REPR=fin, TRANS=vt, вырастить [grow]}
Morph	1262	1269	{MD=imp, NMB=sg, POS=V, REPR=fin, base=вырасти [increase]}
Morph	1270	1275	{POS=ADV, base=вдвое [doubled]}

Table 3. Results of identifying the lookups in the text

Element type	Start position	End position	Attributes
Lookup	0	11	{CAS=acc, GEND=m, NMB=sg, POS=A, base=центральный [central], majorType=locPreKey}
Lookup	0	11	{CAS=nom, GEND=m, NMB=sg, POS=A, base=центральный [central], majorType=locPreKey}
Lookup	12	29	{CAS=acc, GEND=m, NMB=sg, POS=N, base=федеральный округ [federal district], locKind=whole, locSubkind=federal, majorType=locKey, minorType=okrug}
Lookup	12	29	{CAS=nom, GEND=m, NMB=sg, POS=N, base=федеральный округ [federal district], locKind=whole, locSubkind=federal, majorType=locKey, minorType=okrug}
.....			
Lookup	146	158	{CAS=acc, GEND=m, NMB=sg, POS=N, base=Новый Регион [New Region], majorType=SMI}
Lookup	146	158	{CAS=nom, GEND=m, NMB=sg, POS=N, base=Новый Регион [New Region], majorType=SMI}
.....			
Lookup	186	193	{CAS=gen, GEND=f, NMB=sg, POS=N, base=Украина [Ukraine], majorType=location, minorType=republic}
Lookup	186	193	{CAS=acc, GEND=f, NMB=pl, POS=N, base=Украина, majorType=location, minorType=republic}
.....			
Lookup	265	281	{CAS=gen, GEND=m, NMB=sg, POS=N, base=премьер-министр [prime minister], majorType=jobTitle}
Lookup	265	281	{CAS=acc, GEND=m, NMB=sg, POS=N, base=премьер-министр [prime minister], majorType=jobTitle}
.....			
Lookup	340	346	{CAS=dat, COUNTRY=Россия [Russia],

			GEND=f, NMB=sg, POS=N, base=Москва [Moscow], majorType=location, minorType=town}
Lookup	340	346	{CAS=prp, COUNTRY=Россия [Russia], GEND=f, NMB=sg, POS=N, base=Москва [Moscow], majorType=location, minorType=town}
Lookup	350	354	{CAS=gen, GEND=m, NMB=sg, POS=N, base=июнь [June], majorType=time, minorType=month}

.....

Lookup	653	660	{GEND=m, MD=ind, NMB=sg, POS=V, REPR=fin, TNS=past, TRANS=vt, base=пояснить [explain], majorType=speechverb}

.....

Lookup	1182	1185	{CAS=nom, GEND=n, NMB=pl, POS=N, base=МИД [MFA], majorType=gov}
Lookup	1182	1185	{CAS=nom, GEND=m, NMB=sg, POS=N, base= МИД [MFA], majorType=gov}

.....

Lookup	1182	1185	{CAS=dat, GEND=n, NMB=sg, POS=N, base=МИД [MFA], majorType=gov}
Lookup	1207	1214	{GEND=m, MD=ind, NMB=sg, POS=V, REPR=fin, TNS=past, TRANS=vt, base=заявлять [declaire], majorType=announce}
Lookup	1207	1214	{GEND=m, MD=ind, NMB=sg, POS=V, REPR=fin, TNS=past, TRANS=vt, base=заявлять [declaire], majorType=speechverb}
Lookup	1236	1243	{CAS=gen, GEND=f, NMB=sg, POS=N, base=Украина [Ukraine], majorType=location, minorType=country}

Table 4. Results of segmenting the text into the sentences

Element type	Start position	End position	Attributes
Sentence	0	54	{}
Sentence	57	111	{}
Sentence	114	355	{}
Sentence	357	539	{}

Sentence	541	670	{}
Sentence	672	915	{}
Sentence	917	1145	{}
Sentence	1147	1281	{}

Table 5. Results of identifying the named entities in the text

Element type	Start position	End position	Attributes
Location	0	29	{base=федеральный округ [federal district], minorType=okrug}
Date	46	54	{day=26, month=06, year=08}
Person	57	63	{FAMIL=Ющенко [Uschenko]}
.....			
Person	194	207	{FAMIL=Ющенко [Uschenko], FNAME=Виктор [Victor]}
JobTitle	265	281	{base=премьер-министр [prime minister]}
Person	282	296	{FAMIL=Тимошенко [Timoshenlo], FNAME=Юлия [Julia]}
Person	319	337	{FAMIL=Путин [Putin], FNAME=Владимир [Vladimir]}
Location	340	346	{COUNTRY=Россия [Russia], base=Москва [Moscow], minorType=town}
Date	347	354	{day=28, month=июнь [June]}
Person	367	373	{FAMIL=Ющенко [Uschenko]}
Person	375	384	{FAMIL=Тимошенко [Timoshenko]}
Person	408	415	{FAMIL=Путин [Putin]}
Person	663	669	{FAMIL=Ющенко [Uschenko]}
Location	722	729	{base=Украина [Ukraine], minorType=republic}
Location	796	802	{base=Европа [Europe], minorType=area}
JobTitle	897	914	{base=президент [president], jobtitle=президент Украины [Ukrainian president]}
Person	927	933	{FAMIL=Ющенко [Uschenko]}
Person	935	944	{FAMIL=Тимошенко [Timoshenko]}
.....			
SMI	1159	1173	{base=Новый Регион [New Region]}
GovStructures	1187	1197	{$view_name$=МИД России MFA of Russia], Nationality=Россия [Russia]}
Person	1198	1211	{FAMIL=Лавров [Lavrov], FNAME=Сергей [Sergey],

			Status_Role=[глава [head]]}
Location	1241	1248	{base=Украина [Ukraine], minorType=republic}

Table 6. Results of identifying the coferences in the text

Relation type	Attributes
TheSame	{from=15870, to=15867}
TheSame	{from=15866, to=13018}
TheSame	{from=15868, to=15866}
TheSame	{from=15863, to=15871}
TheSame	{from=10861, to=10851}
TheSame	{from=10851, to=10846}
TheSame	{from=15867, to=15864}
TheSame	{from=15869, to=15868}
TheSame	{from=13136, to=13095}
TheSame	{from=15859, to=13018}

Table 7. Results of identifying the semantically meaningful relations in the text

Relation type	Attributes
Personal_and_mediated_contacts	{name=bonds_with_other_persons, from=15861, to=15859}
Personal_and_mediated_contacts	{name=bonds_with_other_persons, from=15860, to=15859}
PresentedIn	{name=contains_inf_on, from=15868, to=8404}
.....	
BeEmployeeOf	{name=works, from=12992, to=15863}
Newsmaking_Person_About	{name=mentions, from=13018, to=15865}
Newsmaking_Person_About	{name=mentions, from=15866, to=13136}
Newsmaking_Person_About	{name=mentions, from=15869, to=15870}
.....	
Newsmaking_Person_About	{name=mentions, from=13018, to=15864}
Newsmaking_Kgether	{name=is_mentioned_with, from=15865, to=13095}
Newsmaking_Kgether	{name=is_mentioned_with, from=15864, to=13095}
Newsmaking_Kgether	{name=is_mentioned_with, from=15870, to=10851}
Newsmaking_Kgether	{name=is_mentioned_with, from=10851, to=15870}
Newsmaking_About	{name=citations, from=13095, to=14566}

Newsmaking_About	{name=citations, from=15864, to=14566}
Newsmaking_About	{name=citations, from=15865, to=14566}
Newsmaking_Kgether	{name=is_mentioned_with, from=15867, to=13136}
Newsmaking_Kgether	{name=is_mentioned_with, from=15864, to=15865}
Personal_Meeting	{name= meetings_and_negotiations, from=15864, to=13095}
Personal_Meeting	{name= meetings_and_negotiations, from=15867, to=13136}
Newsmaking_About	{name=citations, from=15867, to=14568}
Newsmaking_About	{name=citations, from=13136, to=14568}
Newsmaking_About	{name=citations, from=15870, to=14572}
Newsmaking_About	{name=citations, from=10851, to=14572}
BeEmployeeOf	{name=works, должность=head, from=13194, to=15872}
Newsmaking_About	{name=citations, from=10861, to=14574}
Newsmaking	{name=statements_claims, from=13018, to=14566}
Newsmaking	{name=statements_claims, from=15866, to=14568}
Newsmaking	{name=statements_claims, from=15868, to=14570}
Newsmaking	{name=statements_claims, from=15869, to=14572}
Newsmaking	{name=statements_claims, from=13194, to=14574}

Table 8. Triple representation of the text processing results

Object	Attribute	Value
.....		
Новый Регион [New Region]	is_a	СМИ [Mass_Media]
МИД России [Russian Foreign Office]	is_a	Властная_Структура [Commanding_Structure]
МИД России[Russian Foreign Office]	Госпринадлежность [State Affiliation]	Россия [Russia]
Сергей Лавров [Seguey Lavrov]	is_a	Персона [Person]
Сергей Лавров [Seguey Lavrov]	Name	Сергей [Sergey]
Сергей Лавров [Seguey Lavrov]	Инициал [Initial letter]	С. [S]
Сергей Лавров [Seguey Lavrov]	Пол [Gender]	m
Сергей Лавров [Seguey Lavrov]	Статус_Роль [Status_Role]	глава [Head]
Сергей Лавров [Seguey Lavrov]	Фамилия [Family name]	Лавров [Lavrov]
цена на газ для Украины в 2009 году может вырасти вдвое [the price for gas for Ukraine could be doubled]	is_a	Высказывание [Statement]
Украина [Ukraine]	is_a	Местоположение [Location]
.....		
высказывания_заявления [declarations_statements]	is_a	Relation
высказывания_заявления [declarations_statements]	от [from]	479
высказывания_заявления [declarations_statements]	к [to]	740
автор [Author]	is_a	Relation
автор [Author]	от [from]	740
автор [Author]	к [to]	479
упоминает [mentions]	is_a	Relation
упоминает [mentions]	от [from]	479
упоминает [mentions]	к [to]	289
упоминается [is mentioned]	is_a	Relation
упоминается [is mentioned]	от [from]	289
упоминается [is mentioned]	к [to]	479
работает [works]	is_a	Relation

работает [works]	от [from]	479
работает [works]	к [to]	837
работает [works]	Должность [Appointment]	глава [Head]

.....

цитаты [citations]	is_a	Relation
цитаты [citations]	от [from]	289
цитаты [citations]	к [to]	740

Table 9. Fragment of the semantic index of the text (named entities)

Entity (object with attributes)	Semantic index (unique code)
.....	
Person Виктор Ющенко [Victor Uschenko] *First_name* Виктор [Victor] *Initial_letter* В. [V.] *Gender* m *Family_name* Ющенко [Uschenko]	a_7986d80e781ce95b_253defa0906d86a3
.....	
Mass_media Новый Регион [New Region]	a_2d0a16bffbc7dd8f_d77ae55aab932dc5
Commanding_Structure МИД России [MFA of Russia] *State_Affiliation* Россия [Russia]	a_e6eb3a7361bf01ed_68df48c399e93e3e
Person Сергей Лавров [Sergey Lavrov] *First_name* Сергей [Sergey] *Initial_letter* С. [S.] *Gender* m *Status_Role* глава [head] *Family_name* Лавров [Lavrov]	a_2a7aebf65ce54a03_b14a4639a641808d
Statement цена на газ для Украины в 2009 году может вырасти вдвое [prices for gas for Ukraine may be doubled in 2009]	a_f2c3f63347540d89_366c2bf424e720eb
Location Украина [Ukraine]	a_d86dbb670872b609_78aed5dc39a8e2b5
.....	

Table 10. Fragment of the semantic index of the text (pairs of the kind "named entity – named relation")

Pair (object – relation)	Semantic index (unique code)
.....	
Person Сергей Лавров [Sergey Lavrov] *First_name* Сергей [Sergey] *Initial_letter* С. [S.] *Gender* m	b_348eda6c3d9a407_3ebd0c4b59f6bdd0

Status_Role глава [head] *Family_name* Лавров [Lavrov] **Relation** высказывания_заявления [declarations_statements]	
Statement цена на газ для Украины в 2009 году может вырасти вдвое [prices for gas for Ukraine may be doubled in 2009] **Relation** автор [Author]	b_7986d80e781ce95b_253defa0906d86a3
Person Сергей Лавров [Sergey Lavrov] *First_name* Сергей [Sergey] *Initial_letter* C. [S.] *Gender* m *Status_Role* глава [head] *Family_name* Лавров [Lavrov] **Relation** упоминает [mentions]	b_4d188f5919b61bd6_a7d4a195ada52738
Location Украина [Ukraine] **Relation** упоминается [mentioned]	b_beda786d3a1455b4_86ad15ca5e326c5b
Person Сергей Лавров [Sergey Lavrov] *First_name* Сергей [Sergey] *Initial_letter* C. [S.] *Gender* m *Status_Role* глава [head] *Family_name* Лавров [Lavrov] **Relation** работает [works]	b_f6ccf8d7bc4b0ef0_868a50674063a2dc
Location Украина [Ukraine] **Relation** цитаты [citations]	b_47a807f397b5d33b_2ebd0c0ea546446e

.....

Table 11. Fragment of the semantic index of the text
(triples of the kind "named entity – named relation – named entity")

Triple (object – relation – object)	Semantic index (unique code)
.....	
Person Сергей Лавров [Sergey Lavrov] *First_name* Сергей [Sergey] *Initial_letter* C. [S.] *Gender* m *Status_Role* глава [head] *Family_name* Лавров [Lavrov] **Relation** высказывания_заявления [declarations_statements] **Statement** цена на газ для Украины в 2009 году может вырасти вдвое [prices for gas for Ukraine may be doubled in 2009]	c_348eda6c3d9a407_253defa0906d86a3
Statement цена на газ для Украины в 2009 году может вырасти вдвое [prices for gas for Ukraine may be doubled in 2009] **Relation** автор [Au-	c_7986d80e781ce95b_a7d4a195ada52738

thor] **Person** Сергей Лавров [Sergey Lavrov] *First_name* Сергей [Sergey] *Initial_letter* C. [S.] *Gender* m *Status_Role* глава [head] *Family_name* Лавров [Lavrov]	
Person Сергей Лавров *First_name* **Person** Сергей Лавров [Sergey Lavrov] *First_name* Сергей [Sergey] *Initial_letter* C. [S.] *Gender* m *Status_Role* глава [head] *Family_name* Лавров [Lavrov] **Relation** упоминает [mentions] **Location** Украина [Ukraine]	c_4d188f5919b61bd6_86ad15ca5e326c5b
Location Украина [Ukraine] **Relation** упоминается [mentioned] **Person** Сергей Лавров [Sergey Lavrov] *First_name* Сергей [Sergey] *Initial_letter* C. [S.] *Gender* m *Status_Role* глава [head] *Family_name* Лавров [Lavrov]	c_beda786d3a1455b4_2ebd0c0ea546446e
Person Сергей Лавров [Sergey Lavrov] *First_name* Сергей [Sergey] *Initial_letter* C. [S.] *Gender* m *Status_Role* глава [head] *Family_name* Лавров [Lavrov] **Relation** работает [work] **Gov_Structure** МИД России [MFA of Russia] *State_Affiliation* Россия [Russia]	c_f6ccf8d7bc4b0ef0_68df48c399e93e3e
Location Украина [Ukraine] **Relation** цитаты [citations] **Statement** цена на газ для Украины в 2009 году может вырасти вдвое [prices for gas for Ukraine may be doubled in 2009]	c_47a807f397b5d33b_366c2bf424e720eb

.....

Table 12. Fragment of the semantic index of the new text (named entities)

Entity (object with attributes)	Semantic index (unique code)
.....	
Mass_media Новый Регион [New Region]	a_2d0a16bffbc7dd8f_d77ae55aab932dc5
Location Украина [Ukraine]	a_d86dbb670872b609_78aed5dc39a8e2b5
Person Виктор Ющенко [Victor Uschenko] *First_name* Виктор [Victor] *Initial_letter* B. [V.] *Gender* m *Fam-*	a_7986d80e781ce95b_253defa0906d86a3

ily_name Ющенко [Uschenko]	
Person Ангела Меркель [Angela Merkel] *First_name* Ангела [Angela] *Initial_letter* A. [A.] *Gender* f *Family_name* Меркель [Merkel]	a_b5b8f5d8c84d5db1_1fec7102e0d017ab
Персона Виктор Андреевич Ющенко [Victor Andreevich Uschenko] *First_name* Виктор [Victor] *Отчество* Андреевич [Andreevich] *Initial_letter* B. [V.] *Initial_letter2* A. [A.] *Gender* m *Family_name* Ющенко [Uschenko]	a_7cd1659a3047c3bc_6843f8d1284fdcf0

.....

Table 13. Fragment of the semantic index of the new text (pairs of the kind "named entity – named relation")

Pair (object – relation)	Semantic index (unique code)
.....	
Person Виктор Ющенко [Victor Uschenko] *First_name* Виктор [Victor] *Initial_letter* B. [V.] *Gender* m *Family_name* Ющенко [Uschenko] **Relation** встречи_и_переговоры [meetings_and_negotiations]	b_1fec7102e0d017ab_b5b8f5d8c84d5db1
Person Ангела Меркель [Angela Merkel] *First_name* Ангела [Angela] *Initial_letter* A. [A.] *Gender* f *Family_name* Меркель [Merkel] **Relation** встречи_и_переговоры [meetings_and_negotiations]	b_b5b8f5d8c84d5db1_1fec7102e0d017ab

.....

Table 14. Fragment of the semantic index of the new text (triples of the kind "named entity – named relation – named entity")

Triple (object – relation – object)	Semantic index (unique code)
.....	
Person Виктор Ющенко[Victor Uschenko] *First_name* Виктор [Victor] *Initial_letter* B. [V.] *Gender* m *Family_name* Ющенко [Uschenko] **Relation** встречи_и_переговоры [meetings_and_negotiations] **Person** Ангела Меркель [Angela Merkel] *First_name* Ангела [Angela] *Ini-*	c_7986d80e781ce95b_b5b8f5d8c84d5db1

tial_letter A. [A.] *Gender* f *Family_name* Меркель [Merkel]	
Person Ангела Меркель[Angela Merkel] *First_name* Ангела [Angela] *Initial_letter* A. [A.] *Gender* f *Family_name* Меркель [Merkel] **Relation** встречи_и_переговоры [meetings_and_negotiations] **Person** Виктор Ющенко [Victor Uschenko] *First_name* Виктор [Victor] *Initial_letter* B. [V.] *Gender* m *Family_name* Ющенко [Uschenko]	c_b5b8f5d8c84d5db1_7986d80e781ce95b

.....

Table 15. Fragment of the semantic index of the named entities for collection from two texts prior to combining such objects

Entity (object with attributes)	Semantic index (unique code)
Person Виктор Ющенко [Victor Uschenko] *First_name* Виктор [Victor] *Initial_letter* B. [V.] *Gender* m *Family_name* Ющенко [Uschenko]	a_7986d80e781ce95b_253defa0906d86a3
Персона Виктор Андреевич Ющенко [Victor Andreevich Uschenko] *First_name* Виктор [Victor] *Отчество* Андреевич [Andreevich] *Initial_letter* B. [V.] *Initial_letter2* A. [A.] *Gender* m *Family_name* Ющенко [Uschenko]	a_7cd1659a3047c3bc_6843f8d1284fdcf0

.....

Table 16. Fragment of the semantic index of the named entities for collection from two texts after combining such objects

Entity (object with attributes)	Semantic index (unique code)
Персона Виктор Андреевич Ющенко [Victor Andreevich Uschenko] *First_name* Виктор [Victor] *Отчество* Андреевич [Andreevich] *Initial_letter* B. [V.] *Initial_letter2* A. [A.] *Gender* m *Family_name* Ющенко [Uschenko]	a_7986d80e781ce95b_253defa0906d86a3, a_7cd1659a3047c3bc_6843f8d1284fdcf0

.....

Claims

1. A method of computerized semantic indexing of natural language text, which method comprising steps of:

– presenting the text to be indexing in an electronic form for the subsequent computerized treatment;

– segmenting the text in the electronic form into elementary units named hereinafter tokens;

– identifying stable phrases in the text during linguistic analysis;

– forming sentences corresponding to the portions of the text;

– in every sentences having the identified stable phrases, during multi-stage semantic-syntactic analysis by addressing the linguistic and heuristic rules formed in the database in the predetermined linguistic environment, hereinafter referred to as rules, identifying the semantically meaningful objects hereinafter referred to as named entities, and the semantically meaningful relations between the named entities hereinafter referred to as named relations;

– within the text being indexed for every identified named relations relating the certain named entities, forming the set of triples, single first type triple corresponding to the relation established by the named relation between two named entities, each of the set of the second type triples corresponding to a value of particular attribute of one of those entities, and each of the set of the third type triples corresponding to a value of particular attribute of the named relation itself;

– at the set of the formed triples, indexing all named entities related by the named relations separately, all pairs of the kind "named entity – named relation", and all triples of the kind "named entity – named relation – named entity", while taking into account the attributes of respective named entities and/or named relations;

– storing in the database the formed triples and the obtained indices together with the reference to the initial text from which those triples have been formed.

2. The method according to claim 1, wherein said tokens hereinafter referred to as the first level elementary units are selected from the group consisting of: words in the form of the series of letters or letters and hyphens; numbers; punctuation marks; and series of spaces.

3. The method according to claim 1, further comprising a step of forming, for every token being a word, a respective second level elementary units hereinafter referred to as morphs based on the morphological analysis.

4. The method according to claim 1, wherein, during said linguistic analysis at the step of forming sentences, converting the first and/or second elementary units (i.e., tokens and morphs) in every sentence into said stable phrases hereinafter referred to as the third level elementary unit by addressing dictionaries and morphological associations stored in advance in the database.

5. The method according to claim 1, further comprising, during said multistage semantic-syntactic analysis, steps of:

– identifying, in the sentence, said named entities considered as the fourth level elementary units in a set of elementary units of the first, second, and/or third levels;

– forming, using said rules, the morphological attributes for every named entity from the morphological attributes of said elementary units of the second and/or third levels (i.e., morphs and/or stable phrases) which constitute this named entity;

– forming, using said rules, the semantic attributes for every named entity from the semantic attributes of the elementary units of the second and/or third levels which constitute this named entity;

– assigning for every named entity a respective type from the application ontology stored in the database according to the topics of the domain, to which the text being indexed relates;

– storing every named entity in the memory together with the type assigned thereto and morphological and semantic attributes determined thereto.

6. The method according to claim 5, further comprising steps of: determining, for every named entity with type assigned thereto, an anaphoric reference considered as the fifth level elementary unit, and storing said anaphoric reference in the database together with type and attributes of the named entity being the antecedent for that anaphoric reference, as well as with the indication of the co-reference between that named entity and the anaphoric reference thereof;

– said named relations considered as the sixth level elementary units are determined using said rules on the basis of the elementary units of the first, second, third, fourth and/or fifth levels;

– determining, using said rules, for every named relation the morphological attributes from the second level elementary units constituting this named relation;

– determining, using said rules, for every named relation the semantic attributes from the elementary units of the first, second, third and/or fourth levels constituting this relation;

– assigning the respective type to every named relation from application ontology stored in the database according to the topics of the domain, to which the text being indexed relates;

– storing in the memory every named relation together with the type assigned thereto and morphologic and semantic attributes determined thereto.

7. The method according to claim 1, further comprising, prior to storing in the database the formed triples and obtained indices, a step of convolving every group of the objects related by co-reference relations into a single object whose

set of the attributes being the combination of the attributes of all object interrelated by the co-reference relations.

8. A method of computerized semantic indexing of natural language text collection, which method comprising all the steps of the method according to claim 1 in reference to the succeeding text being indexed, after which, during the step of storing in the database the formed triples and the obtained indices of the succeeding text, comprising further steps of: comparing, using the linguistic and heuristic rules in the predetermined linguistic environment, the newly derived named entities and named relations with the named entities and named relations already existed in the database, and in the case of identifying similar named entities and/or named relations, the duplicated information is not stored in the database, and respective named entities and/or named relations are supplemented with references to the succeeding texts where they are present and references to the text fragments within each of succeeding texts from which they are derived.

9. The method according to claim 8, wherein said tokens hereinafter referred to as the first level elementary units are selected from the group consisting of: words in the form of the series of letters or letters and hyphens; numbers; punctuation marks; and series of spaces.

10. The method according to claim 8, further comprising a step of forming, for every token being a word, a respective second level elementary units hereinafter referred to as morphs based on the morphological analysis.

11. The method according to claim 8, wherein, during said linguistic analysis at the step of forming sentences, converting the first and/or second elementary units (i.e., tokens and morphs) in every sentence into said stable phrases hereinafter referred to as the third level elementary unit by addressing dictionaries and morphological associations stored in advance in the database.

12. The method according to claim 8, further comprising, during said multistage semantic-syntactic analysis, steps of:

– identifying, in the sentence, said named entities considered as the fourth level elementary units in a set of elementary units of the first, second, and/or third levels;

– forming, using said rules, the morphological attributes for every named entity from the morphological attributes of said elementary units of the second and/or third levels (i.e., morphs and/or stable phrases) which constitute this named entity;

– forming, using said rules, the semantic attributes for every named entity from the semantic attributes of the elementary units of the second and/or third levels which constitute this named entity;

– assigning for every named entity a respective type from the domain ontology stored in the database according to the topics of the domain, to which the text being indexed relates;

– storing every named entity in the memory together with the type assigned thereto and morphological and semantic attributes determined thereto.

13. The method according to claim 12, further comprising steps of: determining, for every named entity with type assigned thereto, an anaphoric reference considered as the fifth level elementary unit, and storing said anaphoric reference in the database together with type and attributes of the named entity being the antecedent for that anaphoric reference, as well as with the indication of the co-reference between that named entity and the anaphoric reference thereof;

– said named relations considered as the sixth level elementary units are determined using said rules on the basis of the elementary units of the first, second, third, fourth and/or fifth levels;

– determining, using said rules, for every named relation the morphological attributes from the second level elementary units constituting this named relation;

– determining, using said rules, for every named relation the semantic attributes from the elementary units of the first, second, third and/or fourth levels constituting this relation;

– assigning the respective type to every named relation from application ontology stored in the database according to the topics of the domain, to which the text being indexed relates;

– storing in the memory every named relation together with the type assigned thereto and morphologic and semantic attributes determined thereto.

14. The method according to claim 8, further comprising, prior to storing in the database the formed triples and obtained indices, a step of convolving every group of the objects related by co-reference relations into a single object whose set of the attributes being the combination of the attributes of all object interrelated by the co-reference relations.

15. A machine-readable medium intended for direct operation in a computer and comprising a program for carrying out the method according to claim 1.

16. A machine-readable medium intended for direct operation in a computer and comprising a program for carrying out the method according to claim 8.

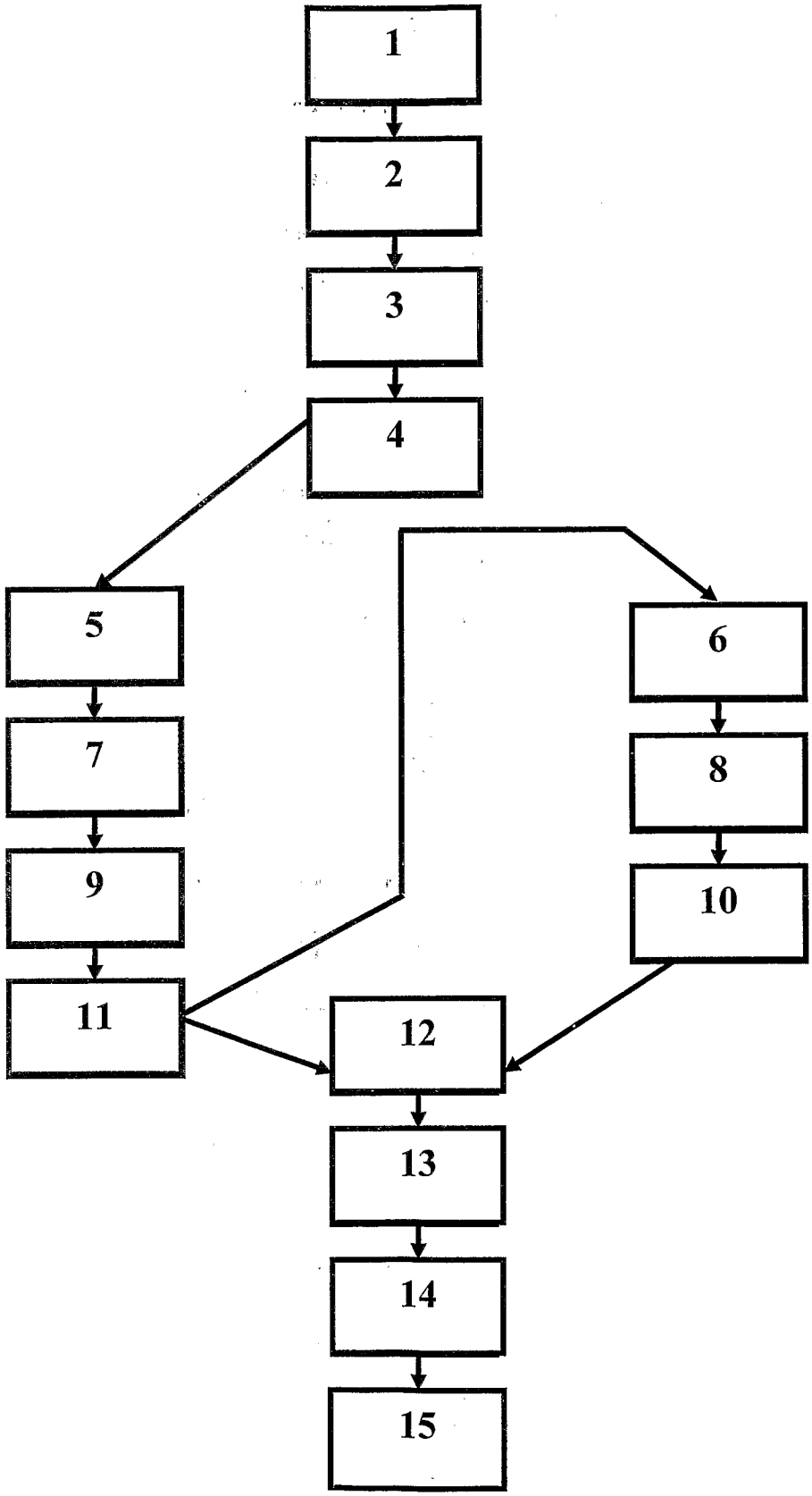

Fig. 1

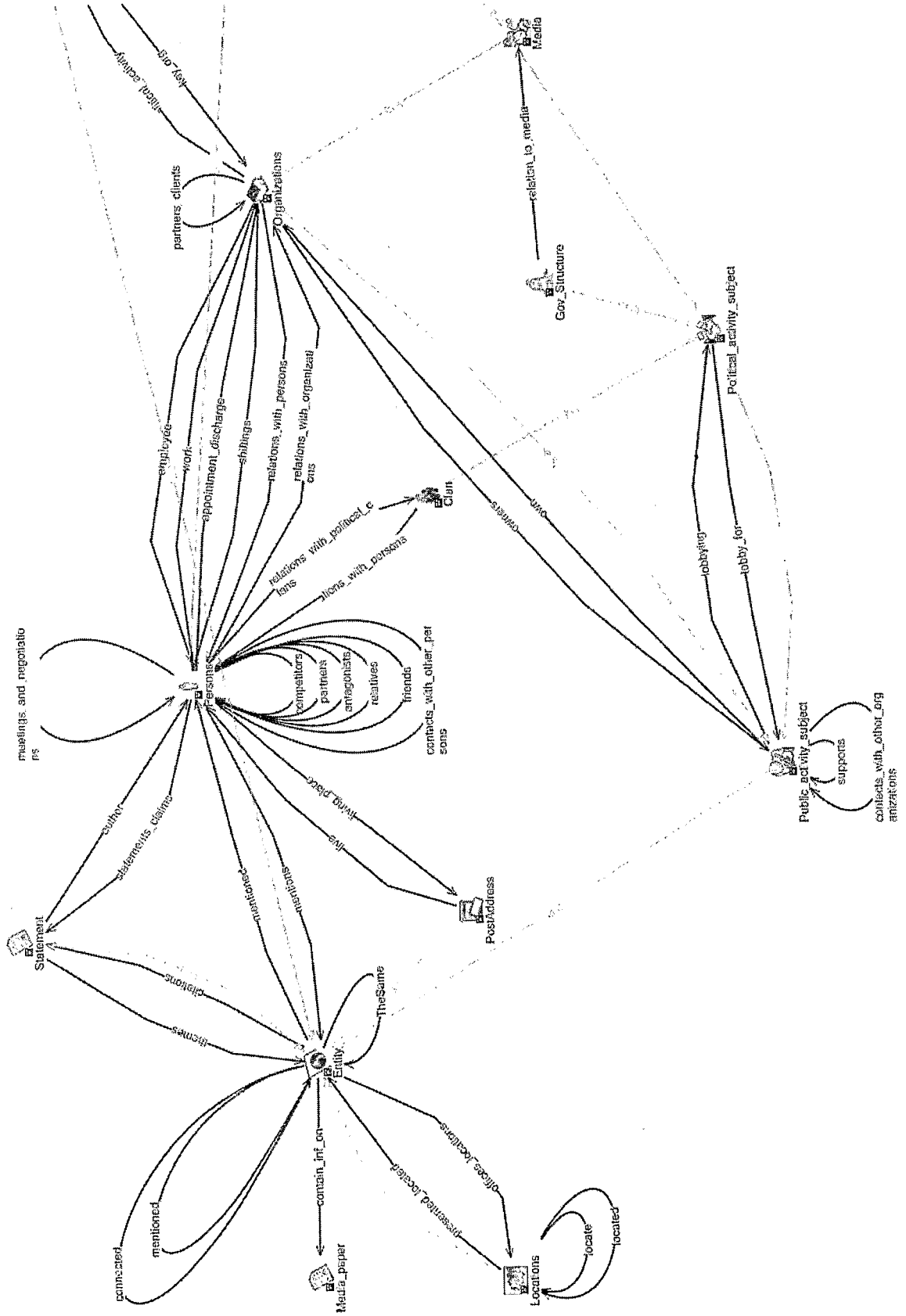
illicit_activity
key_orgs
partners_clients
Organizations
Media
relation_to_media
Gov_Structure
Political_activity_subject
employees
work
appointment_discharge
siblings
relations_with_persons
relations_with_organizations
relations_with_political_figures
Clan
ilions_with_persons
owns
owned
lobbying
lobby_for
meetings_and_negotiations
Persons
competitors
partners
antagonists
relatives
friends
contacts_with_other_persons
living_place
live
author
statements_claims
mentioned
mentioned
Statement
PostAddress
citations
liesIn
TheSame
Entity
Public_activity_subject
supports
contacts_with_other_organizations
contain_inf_on
Media_paper
mentioned
connected
presented_located
offices_locations
Locations
locate
located

Fig. 2

```
(
   {
      (FirstName.GENDER  != null):a,
      (FirstName.NUMB != null):b,
      (FirstName.CASE != null):c
   }
):x +
(
   (
     {
       SecondName.GENDER == :a,
       SecondName.NUMB == :b,
       SecondName.CASE == :c
     }
   ):y
)? +
(
  {
    Morph.GENDER == :a,
    Morph.NUMB == :b,
    Morph.CASE == :c,
    Morph.Orpho == Upperinitial
  }
):z
-->
    Person = {FIRSTNAME= x, SECONDNAME= y, NAME=z}
```

Fig. 3

```
(
  {
      (JobTitle.NUMB != null):a, (JobTitle.CASE != null):b
  }
):x +
(
  {
      Organization.NUMB == sg, Organization.CASE == gen
  }
):y +
(
  {
      Person.NUMB == :a, Person.CASE == :b
  }
):z
-->
    work = {from = z, to = y, JOBTITLE = x}
```

Fig. 4

Как сообщал «Новый Регион», ранее глава МИД России Сергей Лавров заявлял, что цена на газ для Украины в 2009 году может вырасти вдвое.

[As the "New Region" informed, the head of the MFA of Russia Sergey Lavrov told earlier the price for gas for Ukraine in the 2009 could be doubled.]

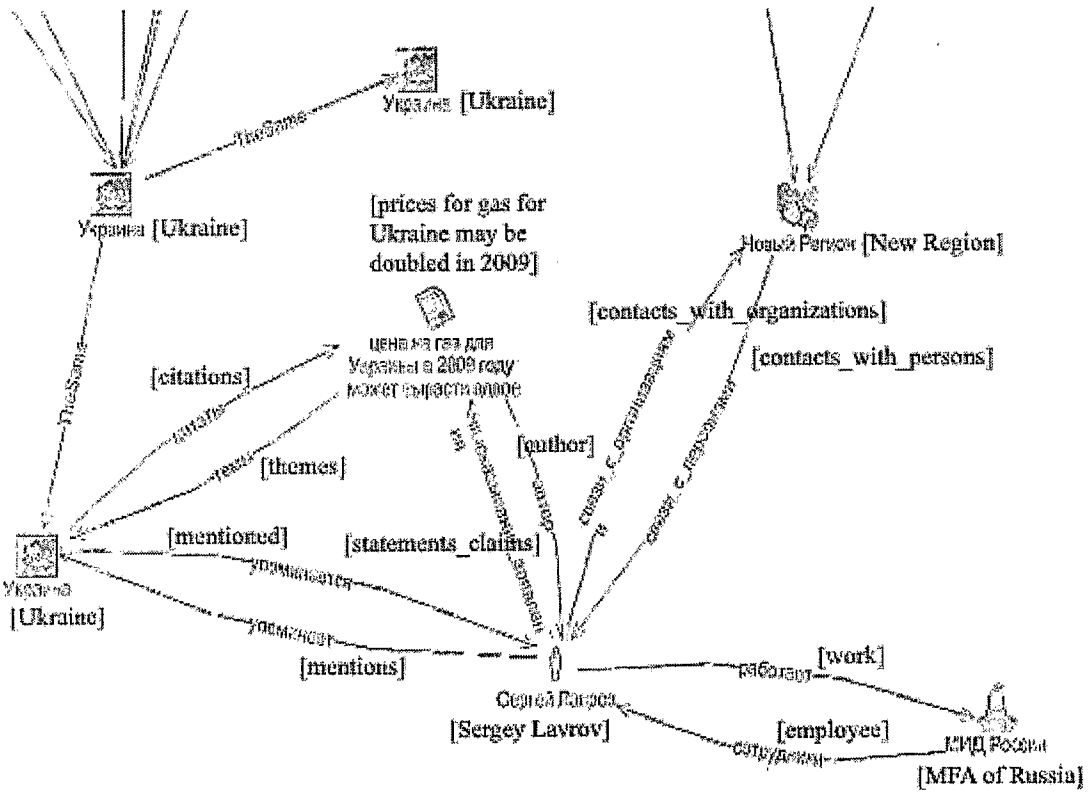

Fig. 5

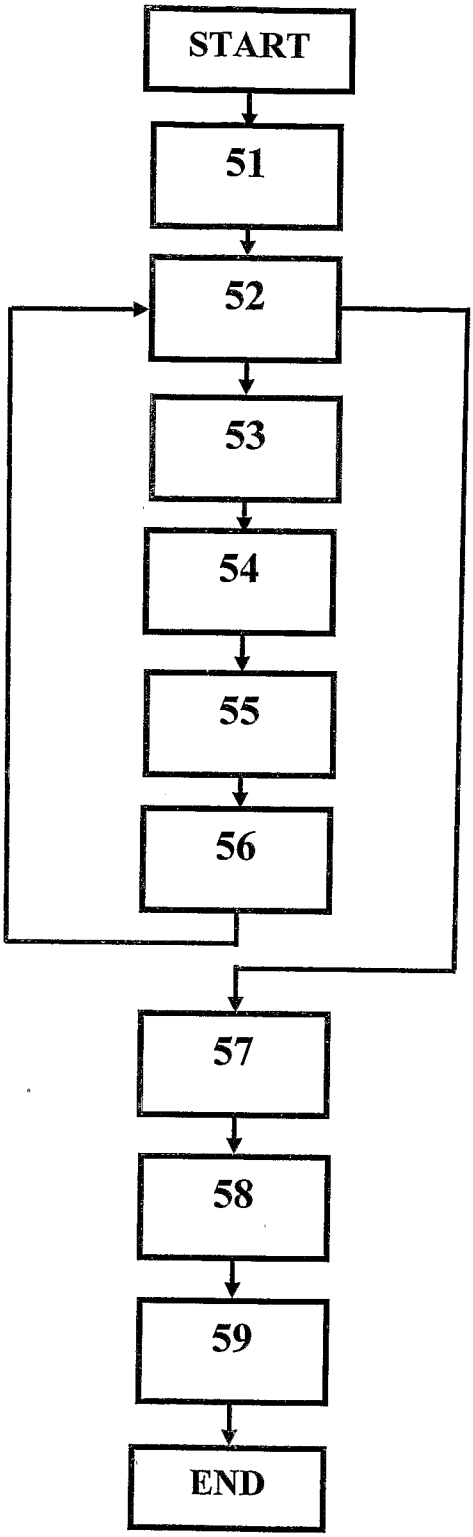

Fig. 6

```
SELECT DISTINCT ?uid ?FamilyName ?FirstName ?EmploymentLabel
WHERE
{
     ?uid
          sofa:__INSTANCEOF_REL onto:Person;
          onto: FamilyName ?FamilyName;
          onto:FirstName ?FirstName.
     ?e
          sofa:__INSTANCEOF_REL onto:work;
          os:_to ?orgUid;
          os:_from ?uid.
     ?orgUid
          sofa:__LABEL_REL ?work.
}ORDER BY ?FamilyName
```

Fig. 7

INTERNATIONAL SEARCH REPORT

International application No.
PCT/RU 2009/000111

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/27 (2006.01)

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G09B 19/00, G09B 19/04-19/08, G06F 17/21, G06F 17/27, G06F 17/28, G06F 17/30

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
PAJ, Esp@cenet, DWPI, PCT Online, USPTO DB, CIPO (Canada PO), SIPO DB, CA (Chem. abstr.), (gallium alloy), PatSearch

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 2007/0073533 A1 (FUJI XEROX CO., LTD.) 29.03.2007, abstract, [0016]-[0078]	1-16
Y	RU 2273879 C2 (NASYPNY VLADIMIR VLADIMIROVICH et al.) 10.04.2006, abstract, claims, p. 26, lines 8-25, p.28, lines 4-18	1-16
A	US 7346493 B2 (MICROSOFT CORPORATION) 18.03.2008	1-16
A	US 7305336 B2 (FUJI XEROX CO., LTD.) 04.12.2007	1-16
A	US 7191115 B2 (MICROSOFT CORPORATION) 13.03.2007	1-16

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is nit considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
09 October 2009 (09.10.2009)	05 November 2009 (05.11.2009)

Name and mailing address of the ISA/RU FIPS	Authorized officer
Russia, 123995, Moscow, G-59, GSP-5, Berezhkovskaya nab., 30-1 Facsimile No. 243-3337	O. Gordeeva Telephone No. (499) 240-2591

Form PCT/ISA/210 (second sheet) (July 2008)