Intelligent Tools for the Semantic Internet Navigator Design

© Igor Kuznetsov © Mikhail Charnine © Elena Kozerenko
© Nikolay Somin © Vladimir Nikolaev © Andrey Matskevich
Institute of Informatics Problems of the Russian Academy of Sciences, Moscow
Igor-kuz@mtu-net.ru keywen1@mail.ru kozerenko@mail.ru
somin@post.ru ipiranlab14@yandex.ru

Abstract

This paper describes the methods and instruments for semantic web navigator design which is a novel system providing semantic drive for Internet users. The solutions proposed rest on the statistical paradigm for knowledge extraction and the semantic presentations based on the Extended Semantic Networks (ESN) mechanism. The approach presented comprises rule-based and stochastic techniques for text processing and extracted entities and relations mapping onto the structures of the knowledge base.

The work is supported by the Russian Foundation for Basic Research, grant 11-06-00476-a

1 Introduction

The paper deals with the issues of design an development of the new tools comprising the intelligent methods and systems based on the presentation mechanism of the extended semantic networks (ESN) [1] which had been employed for creation of a wide range of knowledge-based systems and the features of the keywords encyclopedia Keywen [2].

As a result of tremendous growth of the Internet, its users receive huge volumes of information as responses to their Internet queries. Users are interested in a big variety of questions, they make their own attempts to employ keywords and phrases by means of test and error method (addressing search machines and making analysis of the answers). This results in tremendous expenditures of labour and disappointment because of huge amounts of irrelevant information and/or its incompleteness.

Hence, to make optimal queries, one has to face the problem of requests ordering, reflecting interests of users, creating directories of subjects and articles. It is necessary to create special means, which allow users to find what interests them in the sea of information with reduced expenditures of labour. On-line encyclopedias play the role of such means.

The intelligent web navigator comprises the features of the ESN linguistic processors which were developed for different classes of information systems relating to the artificial intelligence research field.

The core feature of the development is assigning a semantic structure to natural language input. Semantic structure is obtained via the semantic categorization and establishment of semantic relations between concepts presented in natural language texts contained in Internet. *Association* is the dominant type of semantic relations supported by the intelligent tools under discussion. The study of the *contexts* and *co-occurrence* of terms and key words allows to shape the semantic structure of the navigated texts and perform automatic categorization.

The hybrid approach is taken for the semantic navigator design which incorporates the logical analytical functionality of the intelligent systems based on the extended semantic networks, statistical methods and machine learning mechanisms.

2 The ESN Intelligent Systems and their evolution

The intellectual systems, developed on the basis of the apparatus of the extended semantic networks (ESN) [1, 4-6], called the ESN-systems, were created by the association of developers, including the authors of this article at the Institute of Informatics problems of the Russian Academy of Sciences during the period of two decades within the framework of research projects and applied systems, oriented at the concrete subject areas and customers.

We single out 4 generations of ESN- systems. The linguistic semantic ideas laid as the basis of the systems of this class underwent a specific evolutionary process. Intellectual ESN- systems contain the developed bases of knowledge, in this case the knowledge is represented in the form of the records in the language of the extended semantic networks, called ESN - structures. Linguistic knowledge is, thus, a “special case of knowledge” and it is also represented in the form of the

Proceedings of the 14th All-Russian Conference "Digital Libraries: Advanced Methods and Technologies, Digital Collections" — RCDL-2012, Pereslavl Zalesskii, Russia, October 15-18 2012.

records in the language of the extended semantic networks. Basic structural element of the ESN is the named N-ary predicate, called “fragment”. The whole set of language objects are given in the form of predicate-argument structures, in this case the mechanisms for presentation of embedded structures are supported, which gives very powerful presentation mechanisms for describing the objects of different language levels.

The uniformity of language presentations is a very important factor. In the process of analysis and synthesis of natural language sentences the formal grammatical apparatus, similar to the dependency grammar, is used. With this approach the words and the constructions, which perform the role of predicates in the sentence, are the “support” elements, and the result of the analysis of a sentence must become one predicate, which corresponds to the predicate of the sentence (i.e. to basic verb in the tensed form or to another basic predicate expression) in question. Thus, in the process of analysis, in the first place, the processing is performed of the “action words” and the “relation words”, i.e., of the verbs and other words, which have syntactic-semantic valences. An example of a “relation word” the word “father”, “friend”, and the like, i.e., in this case a “relation” is a word which assigns strong clearly expressed syntactical-semantic expectations.

Semantic analysis in the engineering linguistic understanding is the process of translation of natural language expressions into “internal” structures of the knowledge base (KB) in our case these “internal” structures are the records in the ESN language. Thus, a KB structure is the code of sense in the intellectual information systems. The language engineering solutions were implemented in the systems with “complete” linguistic analysis, theses are the systems of the 1st and 2nd generations: DIES1, DIES2, Logos-D [1, 4] and the systems with “factographic” approach, i.e. the intelligent systems of analytical decisions support (ISADS) [6], where the goal of analysis is the extraction of entities and connections from the texts, these are the systems of the 3rd and 4th generations.

The ESN systems of the 4-th generation perform the tasks os semantic objects (named entities) extraction. The set of the objects to be extracted depends on the tasks of a user. At the same time the quality of a linguistic processor is to a considerable degree determined by the possibilities for this extraction. The basic types of information objects and connections, extracted by the ESN semantic processors are given below:

• persons (by family name, given name and patronymic - FNP) with their role features (criminal, victim);
• the verbal description of the persons, their distinctive signs;
• address, posting information attributes;
• date(s) mentioned;
• weapon with its special features;
• telephone numbers, faxes, e-mails with their subsequent standardization;
• the means of transport with the indication of the vehicle type, its state number, color and other attributes;
• passport data and other documents with their attributes;
• explosives and narcotic substances;
• organizations, positions;
• quantitative characteristics (how many persons or other objects participated in an event);
• the numbers of accounts, sums of money with the indication of the currency type;
• terrorist groups and organizations;
• participants of terrorist groups with the indication of their roles (leader, head of, etc.);
• the armed forces, assigned for antiterrorist combat (Military_.Force);
• event (criminal, terrorist, biographical, and so on) with the indication of the information objects participation in them;
• time and the place of events;
• the connection between different types of information objects (with whom a person works in an organization, or lives at the same address, in what events participated together with other objects, etc.). For extracting objects all versions of an object name including the brief form possible in the text were considered. Standard objects (names, dates, addresses, the forms of weapon and others) are reduced to one (standard) form.

The identification of objects is performed taking into account brief designations (for example, separate surnames, patronymics, initials), anaphoric references (indicative and personal pronouns, for example, "this person", "it...") definitions and explanations (for example, "the mayor of Moscow Sobianin" is identified with the subsequent words "mayor", "Sobianin").

For the extraction of events and connections the analysis of verbal forms, participial and adverbial constructions is carried out. An important task is the identification of objects in the entire text, the use for these purposes of indicative pronouns, brief names, anaphoric references.

Taking into account the difficulties and in accordance with the tasks the linguistic processor Semantix was developed, which achieves normalization of words, their grouping with the formation of units, the identification of objects and the establishment of connections. As a result for each NL document a semantic network called the meaningful document portrait was constructed automatically. The latter are the knowledge structures of the knowledge base which serve the basis for implementing different forms of semantic search : the search by features and connections, the search for the objects connected at different levels, the search for similar figurants and incidents, the search by distinctive signs (with the use of ontologies).

The extraction of connections is not only the deep analysis of verbal and other forms. Many connections are given on default. For example, in the summaries of incidents, as a rule, figurants names are followed by their data without the indication of their belonging and with the additional text insertions. For that the directed

search for the connected objects, i.e., the restoration of connections, default data is organized in the processor Semantix.

Special processes are organized in order to connect persons with their place of stay or place of work, vehicles which belong to them, and so forth. For example, the analysis of the summaries of incidents is performed as follows. For a number of objects (address, telephone, date of birth, etc.) a virtual connection with other objects (names, organizations), is built thus yet unidentified. Then, at the same level of processing their search is performed with the aid of the special rules for identification. In these rules the direction of search, the permissible quantity of steps, and also the signs of words and punctuation marks, where the process of search ends are indicated. In this case special filters are required, in order not to take and not to connect an alien object.

This approach showed sufficiently good results in the system Criminal [11]. The special features of natural language are considered where the same actions are identified with the aid of the verbs, verbal nouns and participial constructions. Presented in ESN they are reduced to one form, i.e. a complex object. Moreover, forms with verbal nouns can be the components of verbal forms. On analogy, in ESN some objects can be the components of others. The reason- consequence and temporary dependences between actions, events, etc. are represented which reflect the logical connection of sentences, assigned explicitly, with the aid of the words “therefore”, “then”, etc. The quality of a linguistic processor is determined by a number of factors. First, the possibility for isolation of objects and connections. These are the types of objects being isolated, their quantity. The Semantix processor identifies up to 40 types of objects, including very complex ones, which correspond to actions and events. With an increase in the quantity appear the additional difficulties, connected with collisions of the extraction rules of: some rules can seize the words, which relate to other objects and those extracted by other rules.

It becomes important to consider the order of the application of rules, including of the rules of identification. In the second place, an important factor is the selectivity of rules and procedures of the identification: the factor of the noise and losses. By noise we mean the presence of excessive words in the objects. Losses are the situations when an object is not revealed or revealed partially: in the text there are the words, which did not enter into the object. In the Semantix processor the rules are arranged in such a way that they ensure the high degree of selectivity and the minimization of noise and losses with the large number of the objects being selected.

3 Conceptual linguistic simulation

Conceptual linguistic simulation (CLS) is the process of constructing a natural language model of a subject area (SA) (Fig.1), that synthesizes in itself the approaches of conceptual and linguistic simulation [4-6].

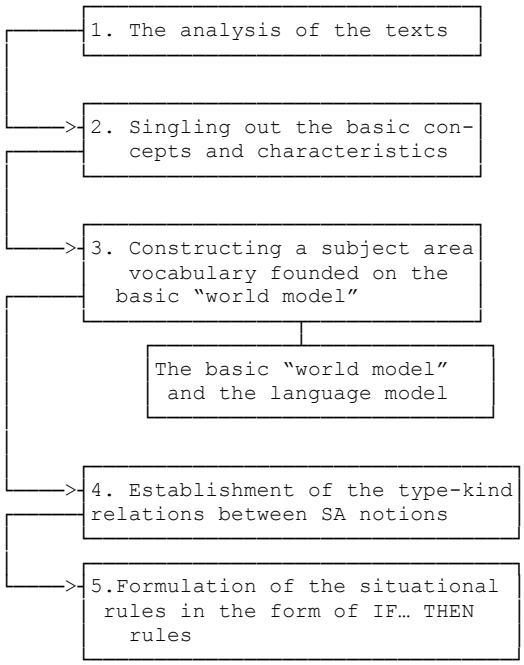

Figure 1 *The flowchart of conceptual linguistic modeling*

Construction of the conceptual linguistic model of a certain subject area is subdivided into the following stages: - construction of the conceptual model proper, i.e., the ramification of fundamental notions, their organization in kind-type trees and the determination of the connections between them; - the development of the ideographic dictionary for the subject area, i.e., the lexical population of the conceptual model; - the introduction of the base rules, which describe "the model of the world" in the natural language relevant for the subject area.

The procedure of conceptual-linguistic simulation on the basis of the ESN apparatus is based on the following principles:

• the model must be "open" , i.e., support the effective mechanism of expansion and information update;

• the model of the “sense” presentation should consider the facts of extra-linguistic reality, which in the form of rules and relations compose a certain basic "world model" and the concrete models of subject areas;

• the model should be practical, i.e., not overloaded by the detailed descriptions of connections and relations between the concepts in order to ensure the possibility of its realization, but at the same time, it should reflect the relevant information for specific objectives.

A realistic approach to the formulation of the problem dictates the need of limitation to a domain-oriented subset of a natural language. The essence of limitations consists in the following: - first, analyzed text materials contain expert knowledge from particular subject areas (we developed the systems for the subject areas for the diagnostics of the microcircuits production failures, forecast in the social sphere, criminology, and others); - in the second place, for the purposes of the

{(ВЫРАБАТЫВА895__)(DICSEM)
COORD(PROGNOZ1,RUS,ВЫРАБАТЫВА895__,S5
0_31_51_20,%) SUB(UNIV,0+) SUB(UNIV,1+)
SUB(UNIV,2+)
ВЫРАБАТЫВ(0-,1-,2-/3+) INFI(3-) ПРИДЕТСЯ(3-)
ПРИДЕТСЯ(3-/4+) FUT1(4-) SUB(СРЕД,5+)

Figure 2 *An example of the presentation of the verb* vyrobatyvat' - "to manufacture" *in the semantic dictionary.*

maximally possible elimination of ambiguity, dictionary is built according to the modular principle: there is a certain most general common part (1-2 levels) completed by special dictionaries for each particular subject area.

The proposed model of lexical semantics is based on the principle of the "nuclear" value realized in the context of this subject area with the subsequent inductive supplementation of other meanings (if they are actualized in the contexts in question). The taxonomy is also used which is realized in the form of the hierarchical trees of the word classes. The general "world model" of the system serves as the basis for the subject area models.

The classes of words, are subdivided into concept/names, relations, actions, properties, characteristics of actions, time and place locatives. The most general notion is "concept", or universal class, which is subdivided into object, the situation, process and others. The words which relate to the classes of actions and relations, are represented as the semantic-syntactic frames, which determine the predicate-argument structures (government model).

However, in the described approach (let us name it the ESN-approach) the range of argument values is substantially extended. This extension consists in the fact that in the role of arguments there can appear simple objects corresponding to the individual words, structural objects which present word combinations, phrases and clauses, and concept of "case" includes not only semantic, but also syntactic aspects. The approach, based on ESN allows to reflect the arbitrary level of the structures embedding it makes it possible to reflect the structural nature of lexical semantics, which in this model has a hierarchical network structure.

Linguistic knowledge is represented in the system dictionary and the declarative modules of linguistic processor. In the ESN systems the function of dynamically formed semantic dictionary which is expanded automatically by the system in the course of concrete texts processing is also realized on the basis of initial linguistic information.

In Fig. 2 the "internal" description of the verb in the semantic dictionary is represented. This dictionary is automatically generated by the ESN-systems DIES2, LOGOS-D, IKS in the course of natural language texts processing.

4 General Considerations for Encyclopedia Design

Encyclopedias traditionally played an important part in the study of new material. However, their creation in electronic type – is a huge work which requires not simply to enter the adequate material into computer, but also and its additional ordering: creation of subject directories for allocation of main classes and subclasses, definition of main notions, building of hyper-references for communication of entries (articles) of encyclopedia between themselves, but also of references to primary sources. What should be also considered is the dynamism of circulating in Internet information: emergence of new information sources, which should be taken into account in encyclopedias.

In Fig. 2 the "internal" description of the verb in the semantic dictionary is represented. This dictionary is automatically generated by the ESN-systems DIES2, LOGOS-D, IKS in the course of natural language texts processing.

At present the majority of large electronic encyclopedias operating on-line have been created on the basis of printed materials of universal encyclopedias: *Big Soviet Encyclopedia*, *Britannica*, *Big Brockhaus*, *Big Larousse* and others. Creation of such encyclopedias requires considerable human labour.

The above said leads us to the conclusion that the global problem in the present situation is the development of methods and program means for automation of the most labor-consuming stages of formation of on-line Internet encyclopedias.

Such formation requires elements of intellectual activity: for making the choice of the subject for description, formation of articles (entries), their names, search for definitions, etc. Development of concepts of on-line encyclopedia results in reference systems of a more general plan, providing collection of information and systematized knowledge representation about different objects which are of interest to the user: - about politicians, persons of science, of culture; - about organizations, companies; - about events (for example, strikes, their reasons, place and time); - about goods and objects of a particular class (for example, fuel, mining, region) and others. While building such systems, many common problems appear, that are also vital for on-line encyclopedia.

The only difference is that instead of articles and their names there would be other objects.

At present the decision of the discussed problems becomes real because there have been designed and developed many systems and facilities in the areas, connected with creating different classes of intelligent systems, language processors, knowledge bases, statistical processing of language components [1-14].

The given work is based on the experience of creation of the on-line encyclopedia and is devoted to the principal directions of decision methods development for the mentioned problem.

5 Special Features of Automation

In general, the problem looks like this. The input comprises a stream of documents from Internet (all relating to a determined application domain). The

output is an electronic encyclopedia consisting of brief articles with names, with hyper-references between articles (if the names of other articles are encountered in the text) and with hyper-references to primary sources the documents from Internet.

In addition an electronic encyclopedia should include the main menu, article sections, various classifiers and the internal search system, providing quick access to concrete subjects making application domain. Certainly, to automate all this processes is not possible.

Formation of the main menu, subjects and query facilities is done manually. Computer can help with selection of material of articles and the choice of their meaningful components.

Two stages are distinguished: training and operation. The grade level, when training sample is given to the system (documents from Internet) with indicated articles which the system should select.

For example, types of diseases can be, symptoms, texts of description, falling into, say, preventive maintenance of diseases and of others. The system should develop decision rules providing allocation of these articles at the stage of operation on other documents.

Such rules are founded on statistical treatment with discovery of keywords and standard contexts (meaningful components), providing selection of articles.

Grade level allows to partly or completely automate the activity of a developer in discovery of the data, necessary for system operation. Discovery of keywords and of contexts requires the use of morphological and semantic blocks of analysis of natural language (NL).

The first block converts word forms

e.g. TABLE, of TABLE, to TABLE

into the uniform type (TABLE) and is particularly important for languages, where words have the a system of cases and other morphological information as, for example, the Russian language.

Without such transformation the search in documents for the same components becomes extremely difficult.

The second block selects word-combinations (they can also be with names of articles) and verbal forms, that determine context in most cases.

Both these blocks of the language processor implementing the analysis of natural language sentences plays an important part in the system.

In creation of on-line encyclopedia important are the following factors: the quality of a created encyclopedia (it is determined by the vicinity to the existing encyclopedia); the difficulty of the preparatory stage including creation and input of basic materials (dictionaries, catalogues and others.) necessary for system operation; also development of a system teaching to discovery of articles is a very difficult programming task.

Simplification of the second and the third factors can dramatically decrease the quality. At the same time, an “over-complication” of the task should be avoided.

We follow the scheme when the development is conducted in stages: first a simple system is developed with subsequent enforcement of its features.

6 Semantic Navigator: Encyclopedia of Keywords

In 2002 the first version of the on-line encyclopedia [2] was released by Michael M. Charnine, having received the name *Encyclopedia of keywords* largely basing on the methods described above. The Encyclopedia functions on the web-site: *www.keywen.com*. It constantly grows and at present contains more than 250000 articles on different subjects in different languages. The majority of the articles are English, but there are also more than 3800 German and 1300 Italian articles. The Encyclopedia of keywords is universally recognized in Internet. Daily several thousand people have free use of its information.

Each article of Encyclopedia consists of key sentences (of phrases). Each of them contains one or several key words. Such phrases are found in Internet with a special semantic navigating program, that is named Keywen Encyclopedia Bot.

At present Encyclopedia contains more than 5 million key-phrases. The major part of the articles of Encyclopedia begin with the section, in which the definitions of terms, included into the article title are given. This allows to understand quickly what the article is about. If a more profound study of the given subject is required, it is possible to use the references to Internet sites. Each phrase is supplied with such reference in Encyclopedia. Each clause of Encyclopedia contains a list of the most important keywords. For each keyword in an article there is a section in which examples of phrases, containing this keyword are given.

The knowledge of keywords is necessary for automatic development of exact requests to search machines. For example, for the article Knowledge Discovery a typical structure in the paragraph DEFINITIONS is given: " Knowledge discovery is the extraction of implicit, previously unknown and potentially useful knowledge from data". An article contains references to more specialized articles: Business and Companies, Magazines and, Organizations, Text Mining, Tools. An article contains keywords (with examples of phrases) KNOWLEDGE DISCOVERY, DATA MINING, INTERNATIONAL CONFERENCE, KDD and others. Encyclopedia (Keywen. com) that contains internal search machine allows to quickly find all key-phrases and appropriate clauses, containing this or that key word. As a result for any keyword it is possible to quickly find application domain corresponding to it. At the beginning of 2004 a version was created of electronic encyclopedia of the Open Project type entitled "Encyclopedia of key phrases". In the framework of this project each user of Internet can bring some contribution into the development of Encyclopedia. The facilities to move sections of any article according to their value and also enter new phrases in Encyclopedia are given to each user.

For Keywen development a constantly growing multilingual texts corpora automatically extracted from Internet is used. For each subject domain and for every supported language a particular text corpus is formed. The text corpora are analysed by the linguistic processor.

Keywen NLP pipeline includes:

a text tokenization module,
a part-of-speech tagging system,
a sentence boundary detection tool,
a collocation identification module,
a named entity recognizer,
a word sense disambiguation system,
a full-syntactic parser.

Extraction of term candidates from domain-oriented texts supports Automatic Term Recognition resulting in Multilingual terminology

Reordering the list of extracted candidates is based on the term/keywords candidate relevance ranking.

Extraction of key phrases and definitions provides Automatic summarization of domain-oriented texts using TF/DF measure

Extraction of key phrases and definitions creates Knowledge-Rich Contexts, automated pattern acquisition is used for the identification of semantic relations: associations and family trees which serve the basis for semantic parser. There are a number of useful advantages of the Keywen apparatus, including, but not limited to: the ability to build large scale human-readable and semantic-oriented hierarchy of categories; the ability to generate dynamical and flexible hierarchical categories; the ability to accept contributions of users with different qualification for improving hierarchical categories; the ability to accept user's minimal contributions (as little as one click); the ability to have multiple ways to categories in the polyhierarchy and at the same time to have hierarchical/directory paths of the categories.

The Keywen apparatus produces a "concrete" substantially repeatable result. It generates hierarchical categories that are substantially repeatable. If users were to perform the claimed steps on multiple different occasions using the same inputs (e.g., the same collection of related terms, the same communication with input/output module), the users would achieve the same result on each occasion. The functionality of the technique has been mathematically proven, the present apparatus for generating hierarchical categories do not use any empiric, heuristic, or fuzzy considerations.

The following two basic category systems are currently most popular:

- Hierarchical, as in directories (easier for understanding, planning and processing);
- Multi-hierarchical, as in Wiki-encyclopedias (more natural, flexible and easy to maintain).

The category structure of Keywen is the product of these two systems: it has advantages of both and opens greater possibilities than either. Both the structure of web-directories and structure of Wiki-encyclopedias may be viewed as an isolated case of Keywen Category Structure. The category structure of Keywen is more precise, logically correct, flexible and dynamic. It is convenient for effective navigation and fast understanding, helps:

- to see the BIG PICTURE,
- to divide knowledge into parts and select the most important parts,
- to create effective plans for learning and knowledge processing.

Hierarchy is a form of organizational structure in which each unit has one and only one "parent" unit, except the "top" unit, which has none.

A Polyhierarchy (multi-hierarchy) is like a hierarchy, but nodes can have multiple parents. In mathematical terms, polyhierarchy is represented by a directed acyclic graph, or a partially ordered set. In terms of object-oriented methodology, it can be viewed as class hierarchy with multiple inheritance.

Directory structure is a particular case of hierarchical structure (that is more general concept). For example, UNIX and DOS have a hierarchical directory structure that allows files to be organized by categories. The main difference between hierarchical and directory structure is different naming convention for categories.

The category names in directory structure can be full or short (local). The full category names in a directory usually are equal to their paths from Top category. A directory contained inside another directory is called a subdirectory of that directory. Subdirectories are specified by concatenating the subdirectory short name to the name of the directory above it in the hierarchy. Together, the directories form a hierarchy, or tree structure.

Keywen Category Structure is a polyhierarchy (multi-hierarchy) that contains one preferred (primary) hierarchy (tree) which contains all nodes.

The following new technologies are employed in Keywen:

- One-click Keywen technology and electronic Voting System,
- Keywen search engine with large queries,
- Keywen Writing Service.

These technologies can accelerate the encyclopedia growth and can make a writer's work most effective.

7 Prospects for the development of Semantic-Focused Systems

The development trends "Encyclopedias of keywords" and "Encyclopedias of key phrases" are determined as follows:

- constant increase of the encyclopedia articles number in different European languages, including Russian, inter-referenced between the relevant articles in different languages;
- the speed of updating of Encyclopedia will be increased; old articles will be kept in the archive of Encyclopedia, but fresh articles will occupy their place with references to the new phrases and new articles from Internet;
- the Rating of articles self-descriptiveness will be

constructed; for this it is necessary to analyze several million references contained in Keywen.com: those containing more key phrases to a given issue, should get high position in the rating.

Further stages of development are connected with the use of language processor.

Stage 1. The system for English and Russian morphological analysis - for transformation of words into normal form. Simplistic analysis of sentences for discovery of definitions on keywords.

Stage 2. The component for analysis of sentences with selection of often met relevant word-combinations.

Stage 3. Means for establishment of relations between relevant objects that form the clauses.

Stage 4. Extension of the notion "meaningful components".

Not only words and word-combinations are allowed, but also objects described in documents: people, addresses, organizations, etc.

Stage 5.Incorporation of the XML-based semantic presentations into the semantic navigator. In the XML file a meaningful portrait of a document (the semantic network structure) is represented comprising all objects and connections, revealed by the Semantix text processor. In connection with this the organization of XML files has the definite scientific value as the means for presentation of the semantic structure of sentences and texts. The transformation of the semantic network into the XML file is ensured with the aid of the reverse linguistic processor. In this case the fragments which present objects, relations, actions and sentences in the semantic network structure are mapped onto the appropriate components of the XML file which will also contain objects, relations, actions and sentences.

The basic task of the LP use consists in operation as a separate module within the framework of the integrated systems of information collection and processing. The exchange is conducted through XML files [14]. For that end a reverse LP was developed, which constructs XML files on the basis of meaningful portraits.

Thus, the input for the linguistic processor (LP) is a natural language text, and the output is an XML- file, where all chosen objects and connections with the indication of sources are represented. This LP named Semantix is provided in the form of an SDK- module. It works under WINDOWS, but it can be recompiled for the work under LINUX.

The Semantix Processor is an independent module and it can be used without the mentioned systems for the standard tasks of analytical services. There are means of tuning to the objects of other types - due to the linguistic knowledge or the dictionaries.

Let us give some explanations. Each object has the following structure:

```
<OBJECT ID="7" TYPE="Organization">
  <ARG CONST="Headquarters />
  <ARG CONST="Residence" />
  <SOURCE> Headquarters residence of the opposing group</SOURCE>
</OBJECT>
```

where ID="7" – is an identification of an object, the TYPE="Organization" is its type. The text component corresponding to the object is also given. Objects relations and their participation in the actions are given through the REF=... references. For example, with the help of the following construction

```
<ACTION ID="15" TYPE="Blow">
  <ARG CONST="At" />
  <ARG REF="7" />
</ACTION>
```

where the sentence "one of the blows struck the headquarters of the oppositional group" is represented. For each object or action the reference to the sentence is given. The Semantix processor uses sufficiently universal constructions of XML- file: one object (through the reference) can include another object. Properties are given as arguments. If necessary the type of attribute is indicated.

For example, in the statement

<ATTR TYPE="YEAR" VALUE="2003"/>

the year is indicated, etc. An XML file has a complete set of information items necessary for the use in different integrated systems.

An example of XML file is given in Figure 3.

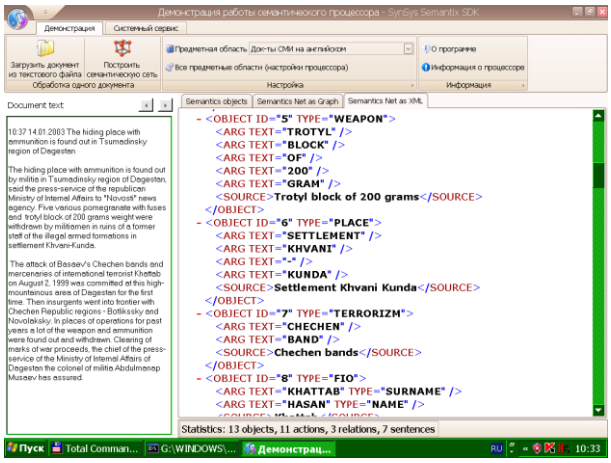

Figure 3. An example of XML file for the semantic structure presentation.

8 Semantic-Focused Systems

The development of concepts of on-line encyclopedia results in more general systems providing discovery of semantically meaningful information from documents, and building on this base an information-reference system [1, 4-6]. The method of tuning - introduction into the system of a new template with the tying of its positions to the components of natural language, or a change in the existing templates and corresponding linguistic knowledge. At present this system is created on the basis of logical-analytical crime detection system ANALYST, and the linguistic processor Semantix using the knowledge base and the semantics- oriented linguistic processor for the tasks of the automatic

formalization of text information, answer to the queries in free form, etc. [4-6].

More than 40 different types of objects are supported by the Semantix processor. The subject areas represented in the text documents are as follows. Documents about terrorism in the Russian language. The analysis of the documents, in which the discussion deals with the terrorist acts and the groups. This feature supports the extraction of 40 types of objects, their connections and the degree of participation in the criminal actions. Documents about terrorists in the English language. The objects and links include persons (their family name, name, patronymic – FNP), posts, organizations, terrorist groups, instruments of crime, time and place of events and so forth, and also connection with and participation in the actions.

- Summaries of incidents. Is ensured the extraction of figurants, their connections, organizations, dates, documents, numbers of bank accounts, details of weapons, etc. with the indication of their participation in particular criminal actions.
- Accusatory conclusions, information about the criminal cases. Objects are identified along the entire field of text. Their connections and criminal actions are revealed.
- Government communications, media issues. Persons, dates, organizations, positions and other significant information and also connections and participation in the actions are selected.
- Autobiographies in the Russian and English languages. From the resumes all attributes of people, periods of time and place of their work, studies, language proficiency and so forth are extracted.
- Autobiographies in the English. From the English language resumes are all attributes of people, periods of time and place of their work, studies, language proficiency and so forth are extracted.
- Documents of media issues in English. From the English language texts the persons mentioned in media issues, positions, organizations, dates, terrorist and anti-terrorist groups, weapons, events, their time and place, different connections and other features are extracted.

In the processors of the Semantix, Lingua-Master, “Criminal” systems up to 40 types of objects are extracted with high accuracy and minimum noise. For example, the system "Criminal" was verified on about 500 thousand incidents from the summaries of Moscow Criminal Police Department, and on the basic objects showed the unique results: the coefficient of noise, i.e. excessive words in the objects) is not more than 1-2% and losses are not more than 1%. The Semantix Processor was fixed on a smaller quantity of documents dealing with the terrorist activity, and therefore there can be more noise and losses in it. But this can be quickly fixed. The fact is that to consider everything which can be encountered in the NL texts is impossible.

Therefore, in the first place, the representative collections of test documents are extremely important, and in the second place, the means of fixing or tuning of linguistic processors are as follows: the employment of hybrid approaches comprising hand-made rules and statistical means for rapid correction and fine adjustment of linguistic knowledge. In our systems there is an entire complex of such means which ensure rapid tuning to the applications (including the introduction of new objects and connections) taking into account the demands of customers.

Such systems have much in common with the system of electronic encyclopedia construction. The significant information corresponds to the names of the articles of the encyclopedia. Templates are the variety of schemes, on which are constructed the articles of the encyclopedia. The layout of the material in accordance with the scheme is required, as well as taxonomic formation of hyper-references.

A backbone instrument for semantic categorization is the employment of hierarchy. Hierarchy is a form of organizational structure in which each unit has one and only one "parent" unit, except the "top" unit, which has none.

A Polyhierarchy (multi-hierarchy) is like a hierarchy, but nodes can have multiple parents. In mathematical terms, polyhierarchy is represented by a directed acyclic graph, or a partially ordered set. In terms of object-oriented methodology, it can be viewed as class hierarchy with multiple inheritance.

Directory structure is a particular case of hierarchical structure (that is more general concept). The main difference between hierarchical and directory structure is different naming convention for categories.

The category names in directory structure can be full or short (local). The full category names in a directory usually are equal to their paths from Top category. A directory contained inside another directory is called a subdirectory of that directory. Subdirectories are specified by concatenating the subdirectory short name to the name of the directory above it in the hierarchy. Together, the directories form a hierarchy, or tree structure.

Keywen Category Structure is a polyhierarchy (multi-hierarchy) that contains one preferred (primary) hierarchy (tree) which contains all nodes.

The method for generating hierarchical categories from collection of related terms contains the following steps:

(a) A huge collection of related terms is accumulated;

(b) Information about relationships of any term is communicated to users (and agents);

(c) Users select multiple parent categories for each term among its relatives;

(d) Many parent-child relationships are accumulated and create direct graph; and

(e) Variety of hierarchical structures is constructed from combined direct graphs of different users.

The last step (e) contains sub steps of:

(e1) Direct graphs of different users are combined together according to user contribution ranks so that better ranked users have the priority in the selection of parents for particular term; and

(e2) Any cycles between nodes in the graph are eliminated.

Categories are indicated in the very beginning of an article; one glance at the category will be sufficient to determine the field of the article, since all categories will contain popular terms.

For example, in the beginning of an article on Mesopotamia, the category "SOCIETY > HISTORY > HISTORICAL ERAS > PREHISTORY > IRON AGE > MESOPOTAMIA" will be indicated.

Even if we do not know the word “MESOPOTAMIA”, the easily understandable words "SOCIETY > HISTORY" will clearly indicate the field.

Categories are located in the beginning of an article; since all categories contain most popular terms, the first glance at the category will make the field of the article clear. All category terms correspond to the titles of the articles, which makes the direction of transition, when mouse-clicking any term within the category, self-explanatory.

Category String is the line that contains the full name of category, which consists of several terms, such as

"THINKING > NONVERBAL THINKING > BIG-PICTURE THINKING".

All terms included into the Category String, are located in hierarchical order, which makes the internal structure of the category easier to understand and more logical. Every category (as full path to category) in Keywen Category Structure is unique.
Keywen Category Structure contains 17 top-level categories.

3.1 ANIMALS > SEA_ANIMALS > WHALES
3.2 ARTS > FILM > ANIMATION > ANIME
3.3 BUSINESS > BUSINESS_ECONOMICS
3.4 COMPUTATION > INTERNET > INTERNET_HISTORY > ARPANET
3.5 GAMES > BOARD_GAMES > KINGS_CRIBBAGE
3.6 HEALTH > MEDICINE > HEALTHCARE > THERAPY > ENERGY_THERAPIES > REIKI
3.7 HOME > COOKING > FRUIT_JUICE > LEMONADE
3.8 IDEAS > BOOKS
3.9 MINERALS > CRYSTALS > ZIRCON
3.10 PEOPLE > POETS
3.11 PLANTS > TREES
3.12 RECREATION > TRAVEL > TOURISM
3.13 REFERENCE > REFERENCE_WORKS > ATLASES > CARTOGRAPHY > WEB_MAPPING
3.14 SCIENCE > NATURAL_SCIENCES > SPACE_SCIENCE > SOLAR_SYSTEM > NEPTUNE
3.15 SOCIETY > HISTORY > HISTORICAL_ERAS > PREHISTORY > IRON_AGE > MESOPOTAMIA
3.16 THINKING > NONVERBAL_THINKING > BIG-PICTURE_THINKING
3.17 WORLD > AFRICA > MIDDLE_EAST > NORTH_AFRICA > EGYPT

9 Conclusions

Thus by semantic navigation we mean semantic analysis and search for the relevant semantic information in natural language texts in the Web. Semantic analysis consists in assigning a semantic structure to natural language input. Semantic structure is obtained via the semantic categorization and establishment of semantic relations between concepts presented in natural language texts. *Association* is the dominant type of semantic relations supported by Keywen and the Navigator under development. Synonyms, taxonomies and other types of paradigmatic semantic relations are established within particular contexts and are viewed as particular cases of the *association* relation. Hence we employ the semantic impacts of *context* and *co-occurrence* which play the decisive role in automatic categorization.

Further development includes the detailed structuring of the Keywen knowledge base with the employment of the Semantix linguistic processor and logical processing features, construction of the encyclopedic articles from definitions and key words automatically extracted from Internet, establishment of hierarchies / category trees on the basis of key word family trees by assigning a dominant category, semi-automatic correction of the category tree, manual and semi-automatic correction of definitions, manual and semi-automatic correction of articles by the methods of digital voting and crowdsourcing. The Keywen technology can be used for terminological data bases creation according to the International Standard ISO 12620: 2009.

The approach taken combines the methods of the rule-based paradigm and machine learning, thus providing a hybrid platform for design and development of the Internet Semantic Navigator.

References

[1] Kuznetsov Igor. Semantic Representations. Moscow: Science, 1986. 294 p. (in Russian)..

[2] Web site for the Keywen encyclopedia of keywords: www.keywen.com

[3] Salton, G. 1989. Automatic text processing: The transformation, analysis, and retrieval of information by computer. New York: Addison-Wesley.

[4] Kuznetsov I., Charnine M. Semantic-Oriented System For Factual Search With the Interface in Russian and English // Systems and Facilities of Informatics. Moscow: Science, 1995, V 7.

[5] Kuznetsov I.P., Efimov D.A., Kozerenko E.B. Tools for Tuning the Semantix Processor to Application Areas // Proceedings of ICAI'09, Vol. I. WORLDCOMP'09, July 13-16, 2009, Las Vegas, Nevada, USA. - CRSEA Press, USA, 2009. P. 467-472.

[6] Kuznetsov I.P., Kozerenko E.B., Kuznetsov K.I., Timonina N.O. Intelligent System for Entities

Extraction (ISEE) from Natural Language Texts // Proceedings of the International Workshop on Conceptual Structures for Extracting Natural Language Semantics - Sense'09, Uta Priss, Galia Angelova (Eds.), at the 17 International Conference on Conceptual Structures (ICCS'09), University Higher School of Economics, Moscow, Russia, 2009. P. 17-25.

[7] Han J., Pei Y. Yin, and Mao R. Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach," // Data Mining and Knowledge Discovery, 8(1), 2004. P. 53–87.

[8] FASTUS: a Cascaded Finite-State Trasducerfor Extracting Information from Natural-Language Text. // AIC, SRI International. Menlo Park. California, 1996.

[9] Cunningham H. Automatic Information Extraction // Encyclopedia of Language and Linguistics, 2cnd ed. Elsevier, 2005.

[10] Dobrov B.V., Lukashevich N.V. Ontologies for natural language processing: Description of concepts and lexical senses // Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference Dialog'06, Bekasovo, May, 31-June, 4, 2006, P. 138-142, 2006.

[11] Kuznetsov I.P., Matskevich A.G. The English Language Version of Automatic Extraction of Meaningful Information from Natural Language Texts // Proceedings of the Dialog-2005 International Conference "Computational Linguistics and Intelligent Technologies", Zvenigorod, 2005pp. 303-311

[12] Web site for Semantic Web: http://www.w3.org/standards/semanticweb/

[13] Jackendoff, R. Semantic Structures. MIT Press, Cambridge, MA, 1990

[14] Gardner, J. R. and Z. L. Rendon, XSLT and XPATH: A Guide to XML Transformations, Prentice Hall, 2001.

Интеллектуальная среда проектирования семантического навигатора по Интернет

И.П. Кузнецов, М.М. Шарнин, Е.Б. Козеренко, Н.В. Сомин, В.Г. Николаев, А.Г. Мацкевич
Институт проблем информатики РАН

В данной статье представлены методы и инструментальные средства создания семантического навигатора по сети Интернет, обеспечивающего новые возможности извлечения семантической информации из текстов, представленных в Интернет на различных естественных языках. Предлагаемые решения основаны на сочетании статистических методов извлечения знаний и механизмов построения семантических представлений на основе аппарата расширенных семантических сетей. Подход, описываемый в статье, включает методы, основанные на правилах и стохастические модели обработки естественно-языковых текстов, а также отображения извлеченных сущностей и отношений в структуры базы знаний.