
Relational-Situational Method for Intelligent Search and Analysis of Scientific Publications

Gennady Osipov
gos@isa.ru

Ivan Smirnov
ivs@isa.ru

Ilya Tikhomirov
tih@isa.ru

Artem Shelmanov
shelmanov@isa.ru

Institute for Systems Analysis of RAS, Moscow, Russia

Abstract

The paper presents model of semantic-syntactic structure of text and method for semantic search and analysis of scientific publications. We mainly focus on stages of semantic analysis of scientific publications and practical usage of developed model and method. Experiments with scientific publication are briefly given.

1 Introduction

Scientific publications normally contain all the essential information about research, including problem statements, proposed solutions and achieved results. Large amount of information depicting the state-of-the-art of science and technology in the world is now openly available on the web. This information can be found in electronic versions of scientific and popular journals, preprints, reports on R&D works presenting the research results, their potential economic impact and recommendations on their usage and applications. However at the moment there is paradox: a scientific papers count constantly grows, but time for paper search and its analysis also grows. For example, PhD-students spend more and more time using digital libraries for papers search and their analysis. Also state-of-the-art understanding in different fields of research or in Science as a whole is a problem in such area, as scientific management. It is important to have a clear view of which research topics are developing at the present time, which topics tend to collapse and which topics are expected to advance in the near future.

This paper presents a model of semantic-syntactic structure of text, method of relational-situational analysis and its practical usage in intelligent search and analytic engine EXACTUS EXPERT. Engine can be used for analytical support of scientific activities.

2 Related work

There are many systems that provide analytical support of scientific activities. However most of them are primarily focused on referential analysis of scientific papers. There are much lesser those that provide abilities for deep comprehensive analysis of scientific trends and domains.

Scopus is a system owned by Elsevier Company. It is positioned as the biggest universal referential database of scientific information. It covers papers from more than 5,000 international publishers including Russian journals. In addition to advanced search ability and referential analysis the system provides tools for comparison of journals by publication activity and by other metrics [MACRVQ⁺07].

Copyright © by the paper's authors. Copying permitted only for private and academic purposes.

In: M. Lupu, M. Salampasis, N. Fuhr, A. Hanbury, B. Larsen, H. Strindberg (eds.): Proceedings of the Integrating IR technologies for Professional Search Workshop, Moscow, Russia, 24-March-2013, published at <http://ceur-ws.org>

SciVal system also proposed by Elsevier is a complex system for science branch analysis. It allows evaluating scientific results of different branches and provides graphical visualization of effectiveness of organizations, countries and geographical regions for a period of time and domain (map of science) [sci].

Web of Knowledge is a system proposed by Thomson Reuters Company. It provides access to the most well-known citation index Web of Science that contains a wide range of publications in almost every branch of science. There are advanced customizable search abilities and tools for referential analysis. The system provides comprehensive statistics about journals and publication for different domains. There are also some tools for monitoring of scientific trends and research teams [LCR13].

Although every system has a lot of abilities and tools, they are not universal but better suit particular cases. The case of scientometrics i.e. determining trends, evaluating research teams and their results, finding novel outstanding approaches is still far away from being solved problem. Another drawback of state-of-the-art systems is that they mostly deal with structured data which is prepared manually. There is a lack of systems that can be fed with raw unstructured data like texts in natural languages. Although there are some examples like Google Scholar that process raw data, these systems provide lesser abilities for analytics of science.

The system we propose EXACTUS EXPERT challenges the tasks of processing the raw unstructured data like texts in natural language and providing tools for comprehensive analysis of scientific paper, domains and trends.

3 Model of semantic-syntactic structure of text

This section is devoted to model that was developed to describe semantic-syntactic structure of text in our experiments. The model is organized into four levels of abstraction: graphematic, morphological, syntactic and semantic.

Graphematic model is the lowest level. In this model text is represented as a hierarchy of elements: sentences, clauses and words. Words are chains of characters produced as a result of tokenization procedure. They consist of letters, numbers, punctuation marks and special characters. Every character in text except whitespace characters belongs to one word. Each word is assigned with a mark that indicates type of the word: abbreviation, punctuation composition and ordinary word.

Clause is a nonempty ordered array of words. Clauses in natural language correspond to simple sentences in a compound sentence, participial phrases, adverbial participial phrases and some other special constructions. Clauses feature projectivity, i.e. they cant overlap partially, but whole clause can be included into another as a part. Finally sentences are nonempty ordered arrays of clauses.

Morphological model extends graphematics model by adding for each word lemma (the canonical form of the word) and some other morphological information. In English morphological information often is represented as a POS (Part-Of-Speech) tag due to limited amount of word forms. On the contrary Russian has a rich, inflexive lexis and lots of different forms what would require hundreds of POS tags. Thus it is common for Russian to consider morphological information of the word as a set of morphological properties which include part of speech as one of its elements. Sets of morphological properties for different parts of speech also differ. For example, nouns feature gender, case, number, when verbs feature tense, personality, number, gender but not the case. In addition we extend noun property set with categorial semantic class (CSC). Categorial semantics is a generalized meaning characterizing words that belong to the same categorial class (for instance, nouns may belong to the classes of people, things and attributes) [OSTZ08]. Categorial semantics class is necessary for semantic analysis since it defines the syntax features of the word and ways of functioning in clause. CSC of word is determined not only by its morphological properties but also by special dictionaries.

The *syntactic* model describes the syntax relations in sentence. The underlying model formalism is dependency tree. There are two types of syntax trees: trees built in a clause connect words and trees built in a sentence connect clauses. The reason to divide syntax relations into two groups is that relations between clauses are more specific than between words. In addition it is easier to build shorter trees than full dependency structure.

For large scale of tasks it is not necessary to build a solid tree, but possible to bring it to number of shallow sub trees. Our algorithms, that perform semantic analysis of text, primarily rely on sub trees that represent NPs (noun phrases) PPs (prepositional phrases) and VBs (verb phrases).

The *semantic* level is represented by relational-situational model [OST10] which is based on the theory of linguistic semantics [OSTZ08], [ZOS04]. The underling relational-situational model formalism is heterogeneous semantic network [ST09], [Osi92].

Nodes of the semantic network are represented by syntaxemes – minimal indivisible semantic-syntactic structures of language [Osi92]. In a particular discourse or in a particular sentence of a query the word acts as a syntaxeme. Syntaxemes are detected according to: a) categorial semantics of the word; b) morphological form; c) function in the sentence.

Syntaxeme that acts as a head of VP (possible NP in case of verbal noun) have special functionality and is called predicate word. In general it holds the central position in the semantic structure of the clause and has influence on related NPs and PPs. Syntaxemes that act as heads of NPs or as main nouns in PPs are called nominal syntaxemes. The nominal syntaxemes bear semantic meanings from predicate words which are represented in the model as labeled relations between nominal syntaxemes and predicate word. There are also semantic relations between nominal syntaxemes that reflect relationships of concepts in the conceptual system of the domain. Semantic relations between nominal syntaxemes express compatibility of their meanings.

The model suggests 65 different meanings of syntaxemes in total, some of them are listed below.

- *Subject* – component that performs action.
- *Ablative* – an initial point of motion.
- *Directive* – direction of motion, oriented action or orientation of object.
- *Mediation* – way (method) of action.
- *Destinative* – appointment of action.
- *Locative* – location of component.

The model suggests 35 different types of semantic relations between nominal syntaxemes, some of them are listed below.

- *ABL* – relationship, where the first component is initial point of motion in direction of the second components destination.
- *TRA* – transitive relationship, where the first component denotes rout of the second component.
- *DIR* – directive relation, where one component denotes the way (direction) of the other component.
- *MED* – mediative relation, whose one component denotes the mode, means of the other component's action.
- *DES* – destinative relation, whose one component denotes destination of the other component.
- *LOC* – relationship where the first component names location of the second component.

Distinct semantics networks of clauses and sentences are linked in semantic network of the whole text by co-referent and anaphoric relations. In addition key concepts in semantic network represented by nominal syntaxemes are enriched with a set of syntactically and semantically related concepts (synonyms, hyponyms, holonyms etc.).

The designed linguistic semantics model allows solving many tasks more effectively in comparison with the known approaches based on keywords extraction [OSTV12]. Semantic network contains the whole essence of text. It allows compiling facts about same object that are mentioned during the whole discourse even if there is a quite a distance between mentions of this object or if it is expressed by different words. It helps to improve relevance of retrieved documents in the information search systems. Semantic networks allow finding documents that are close to the query by meaning. It is also possible to find inferred facts that are not available for search engine if text is represented as a vector of words.

Simplified model of semantic-syntactic structure of text M can be described as follows.

$M = \langle S, T_s, R, I_s \rangle$, where S is a set of syntaxemes $S = \{s_1, s_2, \dots, s_n\}$, s_i – denotes syntaxeme; R – denotes the family of relations on the set of syntaxemes, $R : S \times S$. T_s – denotes syntaxeme types and defined in linguistics dictionaries. $I_s : S \rightarrow T_s$.

Syntaxeme is represented by triple $s = \langle W, P, \tau \rangle$, $\tau \in T_s$, $T_s = \{p, n\}$. Here, W — denotes word; P – denotes syntaxeme features including categorial semantic class, prepositions and other morphological properties; and τ – denotes type of syntaxeme (p – predicate word; n – nominal syntaxeme).

$R = \{(s_1, s_2)\}$ is a family of binary relations on the set of syntaxemes. R consists of three subfamilies: R_p denotes types of relations between predicate words and nominal syntaxemes; R_n denotes types of relations between nominal syntaxemes in a single clause; R_c denotes types of relations that express anaphora and co-reference.

4 Method of Relational-Situational analysis

There are four stages in the semantic processing of the discourse: graphematic, morphological, syntactic and semantic analysis [Zol88], [Osi95]. Each stage is fulfilled by a separate analyzer with its input and output data and its own settings. As the first three stages are quite common only key aspects of stage of semantic analysis will be discussed.

Lets consider text "*Oxygen arrives at tissues from lungs through blood. There it is spent on oxidation of various substances*" as an example. Let us assume that graphematic and morphological structures are already built as well as syntax trees and everything is prepared for the next stage. Note there is only one clause in each sentence so they would be referred in the following discussion as the same components.

The main task of the semantic analysis is to reveal semantic meanings of syntaxemes and relations on a set of syntaxemes. Semantic analysis starts with procedure of predicate word extraction from text. Predicate words mostly are heads of VPs so they are predominantly verbs rarely verbal nouns and participles. In the text above the predicate word of the first sentence is arrive and predicate word of the second sentence is spend. In terms of the model it would be written as follows: $\langle arrive, verb, p \rangle$; $\langle spend, verb, p \rangle$.

Then nominal syntaxemes are extracted from NPs and PPs. There will be four nominal syntaxemes in the first sentence: $\langle oxygen, objective, n \rangle$; $\langle lung, objective;from, n \rangle$; $\langle tissue, objective;at, n \rangle$; $\langle blood, objective;through, n \rangle$ and three in the second sentence: $\langle oxidation, objective;on, n \rangle$; $\langle there, location, n \rangle$; $\langle it, objective, n \rangle$.

When the predicate word and nominal syntaxemes of a clause are determined the nominal syntaxemes are assigned with meanings that correspond to argument structure of the predicate word. Argument structures of predicate word are stored in a special linguistics dictionary. They determine which meanings can be obtained by the syntactically related syntaxemes. The assignment of meanings is mostly based on such features as case (or position in text), categorical semantic class and preposition. Only meanings that correspond to the most completely filled predicate word argument structure are contained. The obtained meanings are stored as a set of relations between predicate word and a nominal syntaxemes. The result set of relations for the first sentence contains four elements: $R_{Subject} = \{(arrive, oxygen)\}$; $R_{Ablative} = \{(arrive, lung)\}$; $R_{Directive} = \{(arrive, tissue)\}$; $R_{Mediation} = \{(arrive, blood)\}$. The result set of relations for the second sentence contains three elements: $R_{Subject} = \{(spend, it)\}$; $R_{Destinative} = \{(spend, oxidation)\}$; $R_{Locative} = \{(spend, there)\}$. In case there is no predicate word in the sentence or it is not found in the dictionary some special algorithms based on machine learning are applied to determine meaning by context.

The next step is to set up relations on the set of nominal syntaxemes. These relations reflect stable relationships between meanings of syntaxemes. Information about computability of meanings is also stored in the linguistics dictionary with the predicate word. The result set of relations of the first sentence contains following relations: $R_{ABL} = \{(oxygen, lung)\}$; $R_{TRA} = \{(lung, tissue)\}$; $R_{MED} = \{(oxygen, blood)\}$; $R_{DIR} = \{(oxygen, tissue)\}$. The result set of relations of the second sentence contains following relations: $R_{DES} = \{(it, there)\}$; $R_{LOC} = \{(it, oxidation)\}$.

After semantic network of each sentence is constructed it is linked in a total semantic network of whole text by co-referent and anaphoric relations. These relations are established using lexical databases like WordNet and features like distance, morphological properties and syntactic role. There are two co-referent relations: $R_{COREF} = \{(it, oxygen); (there, tissue)\}$.

The resulting semantic-syntactic structure of text is represented below.

$$M = \langle S, Ts, R, Is \rangle$$

$$S = \{ \langle arrive, verb, p \rangle; \langle spend, verb, p \rangle; \langle oxygen, objective, n \rangle; \langle lung, objective; from, n \rangle; \langle tissue, objective; at, n \rangle; \langle blood, objective; through, n \rangle; \langle oxidation, objective; on, n \rangle; \langle there, location, n \rangle; \langle it, objective, n \rangle \}$$

$$R_{Subject} = \{(arrive, oxygen); (spend, it)\}; R_{Ablative} = \{(arrive, lung)\}; R_{Directive} = \{(arrive, tissue)\}; R_{Mediation} = \{(arrive, blood)\}; R_{Destinative} = \{(spend, oxidation)\}; R_{Locative} = \{(spend, there)\}$$

$$R_{ABL} = \{(oxygen, lung)\}; R_{TRA} = \{(lung, tissue)\}; R_{MED} = \{(oxygen, blood)\}; R_{DIR} = \{(oxygen, tissue)\}; R_{DES} = \{(it, there)\}; R_{LOC} = \{(it, oxidation)\}; R_{COREF} = \{(it, oxygen); (there, tissue)\}$$

Figure 1 shows visual representation of the built semantic-syntactic structure of text as a semantic network. Syntaxemes are represented by nodes of the semantic network. The solid edges denote relations between predicate words and nominal syntaxemes and they are marked with corresponding meanings. The dashed edges denote relations between nominal syntaxemes and they are marked with corresponding relation types. Finally the dotted edges denote syntax relations and the wide dotted edges with *COREF* mark denote co-referent relations.

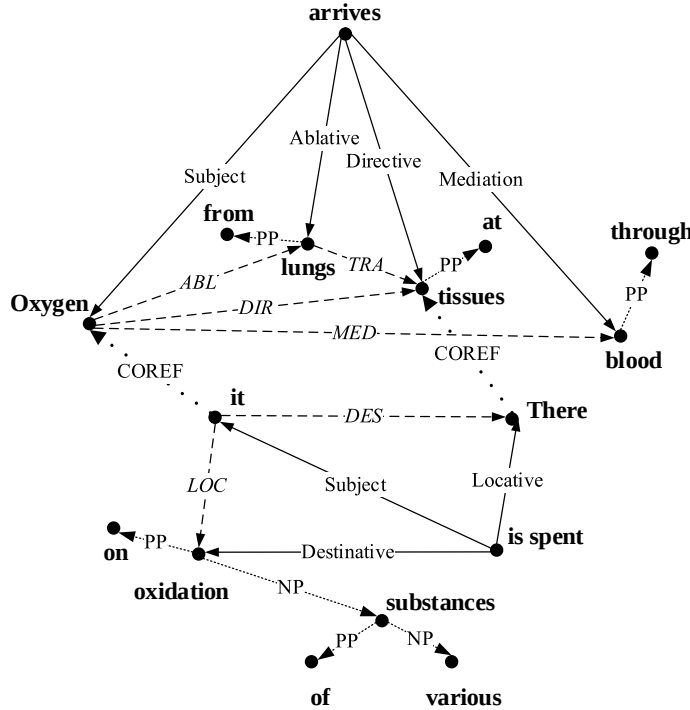


Figure 1: Visual representation of the semantic-syntactic structure of text: "Oxygen arrives at tissues from lungs through blood. There it is spent on oxidation of various substances."

5 Practical usage

Semantic images of scientific text generated as the result of the semantic analysis are stored in semantic indexes and used for search and analysis. The model of semantic-syntactic structure of text described above extends possibilities of text processing and gives additional information about content of scientific publications. It allows extracting, for example, objects of research, methods and tools applied in research, results of research and other useful entities from publications. The solution of following tasks is based on the results of semantic analysis of scientific publications.

5.1 Semantic search

Semantic search of publications allows using queries formulated in natural language. The main idea of semantic search is semantic matching of a query with documents stored in search index. Semantic search involves generation of semantic images of documents and queries. The semantic image as described above is presented as the semantic network so the semantic matching consists in comparison of networks for query and documents meaning by meaning and relation by relation. In the result the semantic relevance is calculated that allows ranging the documents by semantically correspondence to the search query.

Semantic search involves both statistical approaches based on TFIDF and semantic analysis. The semantic analysis substantially enhances search precision and recall and reduces the number of irrelevant documents

returned by the search engine.

5.2 Search for similar documents

The same idea as for semantic search is used for searching similar documents. In this case semantic images of two documents are matched and semantic distance between them is calculated. It allows detection of possible duplicates of scientific papers and plagiarism, tracking of succession (or revelation of its absence) in the results of research work in various types of scientific information sources (R&D reports, technical documentation, research papers, publications in mass media).

5.3 Extraction of definitions

Definitions of terms in scientific publications can be determined by their lexical, syntactic and semantic contexts. To find and extract term definitions we developed method based on analysis of these contexts [She12]. Our linguists revealed more than 60 contexts of term definitions. They were refined and summarized during some experiments. In a result we created set of syntactic and semantic templates that covers the most frequent cases of term definitions. The idea of the method is to search all matches that suit lexical, syntactic and semantic conditions of stored templates. For example, template *POS(Noun) & SemValue(Estimative) + Lemma(called)* matches definition "This dividing line is called the bissectrice or bisection line". Fifteen templates of such kind were implemented.

Templates also bear information about, which part of the match should be extracted as a term and which part should be treated as a definition. When list of terms is constructed it is filtered using some heuristic rules. These rules exclude from the result set typical erroneous definitions and redundant words from terms themselves.

The set of found terms of a document can extend a list of keywords, it can be taken into account by procedure of automatic annotation construction and it also can be a sign of novelty of scientific paper. Terms and definitions can also obtain a special mark in a search index which has influence on relevance of the document to a query. Terms and definitions placed into a search index can help to trace relationships between documents since it is possible to find texts with similar terminology and even determine in what text term was introduced first.

5.4 Retrieving results of research from papers

The results of research presented in the paper are formulated by means of specific phrases which correspond to special structures. We suggested these structures containing pairs $\langle \textit{predicateword}, \textit{meaning} \rangle$ of special predicate word and meaning of its argument (syntaxeme).

For extraction of such structures a corpus of scientific texts with marked up phrases describing results was formed. Using Bayesian classifier structures for extracting results were obtained. For example, a result can be presented with the structure $\langle \textit{develop}, \textit{object} \rangle$, so the sentence "Authors developed the method" is considered describing result.

It was discovered that theoretical results are commonly presented with structures $\langle \textit{predicateword}, \textit{deliberative} \rangle$, and applied results are presented with structures $\langle \textit{predicateword}, \textit{object} \rangle$.

Retrieving results allows evaluating efficiency of a given research or a given field of research and makes possible to compare them by productivity.

5.5 Assessment of the quality of scientific publications

The problem of evaluating quality of scientific publication has two aspects. First, the publication should have conventional format, i.e. contain sections such as problem statement, methods, solutions, results of the research, conclusions, references and so on (see, for example, IMRAD [Day89],[SP04]). Second, the publication should not contain quasi-scientific or prescientific lexis and phrases.

To check the paper's structure it is necessary to detect availability of mentioned sections. As for retrieving results we assume that sections contain specified semantic structures such as $\langle \textit{predicateword}, \textit{argument}, \textit{meaning} \rangle$. A corpus of scientific texts with marked up sections was created. Bayesian method extracted structures specific to the sections. Thus, the section "problem statement" frequently contains structures $\langle \textit{is}, \textit{research}, \textit{object} \rangle$, $\langle \textit{attract}, \textit{attention}, \textit{subject} \rangle$ etc., the section "conclusion" contains structures $\langle \textit{discover}, \textit{opportunities}, \textit{resultative} \rangle$, $\langle \textit{present}, \textit{we}, \textit{subject} \rangle$, $\langle \textit{let}, \textit{research}, \textit{causative} \rangle$ etc.

For checking quasi-scientific or prescientific lexis and phrases in publications the special dictionaries were developed.

6 Experiments results

The described principles, models and methods were implemented in the intelligent search and analytic engine EXACTUS EXPERT.

To evaluate quality of developed semantic analyzer the small corpus of sentences was created. It contains two hundred sentences that consist of single clause, which represent a search queries. The precision on this corpus is 0.83 and the recall is 0.97. The achieved result is good for big data approaches and is suitable for our search engine since it deals with large scale collections of textual documents.

The search algorithm of EXACTUS EXPERT was tested on Russian Information Retrieval Evaluation Seminar (ROMIP)[NN08], a competition of Russian search engines. In many respects ROMIP seminars are similar to other world information retrieval events such as TREC [tre], CLEF, NTCIR, etc. Similar to TREC, ROMIP has cycle nature and is overseen by a program committee consisting of representatives from academia and industry. Several tracks that correspond to different tasks are conducted. In few years on ROMIP there was conducted search in large scale collections task. For example, in 2008 there were two collections containing 1.6 and 3.0 million Russian documents. The evaluation procedure changes from year to year. In general competitors provide search results for a big set of queries of different types (e.g. about 30,000 queries in 2008), which are compared against the ground truth. The ground truth consists of the smaller set of queries (e.g. about 500 queries in 2008) randomly chosen from the big set and assessed by some experts. Several widely known evaluation metrics are used in ROMIP: precision, recall, 11-point TREC precision-recall graph, Bpref etc.

In 2008 our search algorithm showed the highest precision/recall values, in 2009 the algorithm for searching similar documents showed one of the best results. The experiments show the advantage of using linguistic methods together with statistical methods for improvement of search quality.

The experimental study of the developed methods for semantic analysis of scientific publications was carried out on the material of scientific journals, conference proceedings and theses abstracts. We have processed about 100 thousand publications in total, including journal publications in Russian and English, theses abstracts in Russian and conference papers in Russian and English.

The algorithm for retrieving results of research from papers showed value 0.85 for precision on testing data with value 0.90 for precision of detecting theoretical or applied results. The precision of definition and term extraction algorithm is 0.84 and the recall is 0.86 [She12].

Example of report generated by system for a paper quality evaluation is presented below.

Science index (from -5 to 5) equals 4.

Interpretation:

Text contains 28% of scientific lexis and 1% of quasi-scientific or prescientific lexis. References are present. Problem statement is available with probability 0.83. Methods described with probability 0.87. Conclusions are available with probability 0.55.

Results:

Methodological approach described in the paper ... was developed for accessing large distributed informational systems ... and applied for disaster recovery....

Definitions:

Disaster recovery capability to recover applications and data after disaster...

The process of creation of such model is called as disaster recovery modeling...

In the report science index of 4 means that the analyzed paper is more likely to be scientific, because it contains big amount of scientific lexis and lesser amount of quasi-scientific lexis. There is also high probability that paper has a problem statement and a methods description. The system is not sure about presence of conclusion.

7 Conclusion and future work

The developed model of semantic-syntactic structure of text and method help to solve set of tasks of analytical support of scientific activities. The search and analytic engine EXACTUS EXPERT is demanded by experts to support the decision making process on financing of research topics, by editors of scientific journals and by researchers themselves, especially by PhD-students.

As a future work, we plan to develop methods for detection of logical defects in scientific texts. This feature is demanded by editors of scientific journals. Also, we plan to improve quality of English text analysis and conduct more experiments.

7.0.1 Acknowledgements

The project is supported by Ministry of Education and Science of the Russian Federation grant 07.551.11.4003.

References

- [Day89] R. A. Day. The origins of the scientific paper: The imrad format. *American Medical Writers Association Journal*, 4(2), 1989.
- [LCR13] Loet Leydesdorff, Stephen Carley, and Ismael Rafols. Global maps of science based on the new web-of-science categories. *Scientometrics*, 94:589–593, 2013.
- [MACRVQ⁺07] Felix Moya-Anegon, Zaida Chinchilla-Rodriguez, Benjamin Vargas-Quesada, Elena Corera-Alvarez, FranciscoJose Munoz-Fernandez, Antonio Gonzalez-Molina, and Victor Herrero-Solana. Coverage analysis of scopus: A journal metric approach. *Scientometrics*, 73:53–78, 2007.
- [NN08] Marina Nekrestyanova and Igor Nekrestyanov. Romip 2008 evaluation: Rules, methodology and adhoc decisions. In *Proceedings of ROMIP'2008*, pages 5–26, 2008.
- [Osi92] Gennady Osipov. Formulation of subject domain models: Part 1. heterogeneous semantic nets. *Journal of Computer and Systems Sciences International. Scripta Technica Inc.*, 1992.
- [Osi95] G. Osipov. Methods for extracting semantic types of natural language statements from texts. In *10th IEEE International Symposium on Intelligent Control*, Monterey, California, USA, aug 1995.
- [OST10] G. S. Osipov, I. V. Smirnov, and I. A. Tikhomirov. Relational-situational method for text search and analysis and its applications. *Scientific and Technical Information Processing*, 37(6):432–437, 2010.
- [OSTV12] Gennady Osipov, Ivan Smirnov, Ilya Tikhomirov, and Olga Vybornova. Technologies for semantic analysis of scientific publications. In R. R. Yager, V. Sgurev, and M. Hadjiski, editors, *Proceedings of 2012 IEEE 6th International Conference Intelligent Systems*, volume 2, pages 58–62, 2012.
- [OSTZ08] Gennady Osipov, Ivan Smirnov, Ilya Tikhomirov, and Olga Zavjalova. Application of linguistic knowledge to search precision improvement. In *Proceedings of 4th International IEEE conference on Intelligent Systems*, volume 2, pages 17–2–17–5, 2008.
- [sci] Scival <http://info.scival.com/>.
- [She12] A. O. Shelmanov. Method for automatic extraction of multiword terms from texts of scientific publications. In *Proceedings of thirteenth National conference on Artificial Intelligence with international participation CAI-2012*, volume 1, pages 268–274, Belgorod, 2012. BGTU.
- [SP04] L. B. Sollaci and M. G. Pereira. The introduction, methods, results, and discussion (imrad) structure: a fifty-year survey. *J. Med. Libr. Assoc.*, 92(3), 2004.
- [ST09] I. Smirnov and I. Tikhomirov. Heterogeneous semantic networks for text representation in intelligent search engine EXACTUS. In *Proceedings of workshop SENSE'09 - conceptual Structures for Extracting Natural language SEMantics, The 17th International Conference on Conceptual Structures (ICCS'09)*, pages 1–9, Moscow, Russia, July 2009.
- [tre] Trec: Text retrieval conference. <http://trec.nist.gov/>.
- [Zol88] G. A. Zolotova. *Syntactic dictionary: Repertory of elementary units of Russian Syntax*. Nauka, Moscow, 1988.
- [ZOS04] G. Zolotova, N. Onipenko, and M. Sidorova. *Communicative grammar of Russian language*. oscow, 2004.