Government of Russian Federation

Federal state autonomous educational institution of higher professional education

National Research University
Higher School of Economics (HSE)

Syllabus of the course
Semantic Information Systems
for the students of the 4th year of the direction 080500.62 “Business Informatics”
Bachelor program “Business Informatics”

Author: Professor, Ph.D., Dr.Sci. V. A. Fomichov (vfomichov@hse.ru)

Recommended by the section of EMC
“Business Informatics”

Approved at the meeting of the
Department of innovations and
business in the sphere of informational technologies

Chair
__________________________ J. V. Taratukhina
“21” December 2013

Department chair
__________________________ S. V. Maltseva
“07“ November 2013

Moscow - 2013
1 Course Description

a. Title of the Course: Semantic Information Systems

b. Pre-requisites: knowledge of basic ideas of high school algebra, basic notions of set theory, theory of graphs, theory of relational databases, the skills of developing rather simple algorithms, the skills of systematizing and generalizing information. Knowledge of basic ideas of the courses “Discrete Mathematics”, “Programming”, “Web-programming” is desirable.

c. Course type: elective

d. Abstract:

Section 1 of the course describes basic concepts of the field and gives an outline of the principal application fields. In particular, the notions of an applied intelligent system, natural language (NL), a natural language text (NL-text), a semantic representation of a NL-text, ontology, semantic information system, semantic mashup, Semantic Web, Linked Open Data (LOD), and Multilingual Semantic Web are introduced.

The subject of Section 2 is the principal ideas of applying first order logic to representing in a formal way semantic structure of natural language sentences and discourses. The connections between first order logic and the theory of semantic nets and frames are considered.

Section 3 describes the principal ideas of basic informational languages of the Semantic Web project: RDF (Resource Description Framework), RDFS, and OWL (Ontology Web Language). The way of representing triplets (main data structure of RDF) as simple semantic nets is shown. The main ways of determining classes of objects by means of the language OWL are considered. The state of the studies aimed at developing OWL-based ontologies is outlined. The conception of Linked Open Data is discussed. The principal features of the query language SPARQL (for interaction with LOD) are described.

Section 4 sets forth the basic ideas of describing in a formal way semantic structures of sentences and discourses in NL by means of the theory of K-representations (knowledge representations). It is an original theory of designing semantic-syntactic analyzers of NL-texts with broad use of formal means for representing input, intermediary, and output data. The current configuration of this theory is represented in the V. A. Fomichov’s monograph published by Springer in 2010.

The theory of K-representations proposes (for the first time in the world) a system of ten operations on conceptual structures providing the possibility to construct, step by step, semantic representations of arbitrary complex sentences and discourses in English, Russian, German, French and other natural languages. The mathematical model describing this system determines, in particular, a new class of formal languages called SK-languages (standard knowledge languages). The most general ideas of this model are described. The principles of constructing semantic representations (in other terms, text meaning representations) of questions of many kinds, statements, and commands, and complex discourses with the help of SK-languages are set forth. The texts in considered examples pertain, in particular, to business, management, medicine, biology, technology, sport. The advantages of the theory of K-representations in comparison with semantic nets theory, theory of conceptual graphs, first order logic, and episodic logic are analyzed.

Section 5 describes the main available approaches to describing logical structure of linguistic databases, i.e. of databases containing the information used by the algorithms of syntactic analysis and semantic-syntactic analysis for constructing semantic representations of NL-texts.

Section 6 sets forth the principal approaches to semantic-syntactic analysis of NL-texts, in particular, the approaches using ontology for analysis of NL-texts. The connections with the results obtained under the framework of Semantic Web project are described. A considerable attention is paid to the approach of the theory of K-representations to semantic-syntactic analysis of NL-texts. This approach uses original formal means (being understandable for programmers) for describing the algorithms of semantic-syntactic analysis independently on any programming environment. That is why the use of this approach provides good preconditions for transporting the algorithms from one application domain to other domains and makes easier the adaptation of the algorithms to new applications.
Section 7 describes the principal ideas of designing semantics-oriented NL processing systems (NLPSs) in business informatics, biomedical informatics, text mining, and other fields. In particular, the recommender systems with NL-interfaces and the NLPSs for constructing the profile of a person or a company, proceeding from the data dispersed in numerous textual sources, are considered. Besides, this section includes an overview of most interesting projects of NLPSs in the quickly progressing field of bioinformatics. On the one hand, the approaches to designing NL-interfaces to databases with medical and biological information are analyzed. On the other hand, the projects aimed at finding in the scientific literature the descriptions of special experiments with proteins or at discovering the fragments describing the experience of using in practice certain new medicines or procedures are considered.

The final part of Section 7 considers the problem of and the approaches to developing a Semantic Web of a new generation, or Multilingual Semantic Web, or Meanings Understanding Web. Its principal distinguished feature is to be the well developed computer tools for understanding written texts and oral speech in many natural languages: English, Russian, etc. An original strategy of transforming the existing Web into a Semantic Web of a new generation is set forth; this strategy was proposed by V.A. Fomichov in [7, 11, 13, 15]. The central idea of this strategy is the use of a comprehensive semantic formal environment (the class of SK-languages) for represented the conceptual structures (or the meanings) associated with various Web-based sources: natural language texts, visual images, films, the messages sent by computer intelligent agents in multi-agent systems, semantic annotations of Web-based sources, etc.

e. Field of application and normative references

This program of an academic discipline formulates minimal demands to knowledge and skills of students and determines the content and kinds of lessons and control data.

The program is for lecturers, tutors, teaching assistants, and students of the direction 080500.62 “Business Informatics” of bachelors’ preparation studying in accordance with the Bachelor Program “Electronic Business”.

The program is developed in accordance with:
* The educational standard of the federal state autonomous educational institution of higher professional education National Research University Higher School of Economics, the level of preparation: bachelor, confirmed 29.06.2012;
* working academic university plan corresponding to the direction 080500.62 “Business Informatics” of the bachelors’ preparation for the Bachelor Program “ Electronic Business”, confirmed in 2012.

2 Learning Objectives

The learning objectives are to acquire a collection of theoretical knowledge and methodological foundations in the filed of semantic information systems (SIS) and also the practical skills being necessary for developing computer implementations of SIS.

3 Learning Outcomes

As a result of mastering this discipline, a student is:

- To know the characteristics of the state of the art of the field of design and application of semantic information systems (SIS), the development trends of this field, the principal classes of SIS, the principles of designing main subsystems of SIS, the main approaches to formalizing semantic structure of sentences and discourses in natural language (NL);
- To be able to represent in a formal way the descriptions of various classes of entities (objects) by means of the Semantic Web project languages RDF, RDFS, OWL;
- To be able to represent in a formal way the semantic content of NL-texts by means of SK-languages (standard knowledge languages), determined by the theory of K-representations;
- To possess the skills of reflecting in a formal way the semantics of lexical units (the words and short word combinations) with the help of SK-languages.
• As a result of mastering this discipline the student acquires the following competencies:

<table>
<thead>
<tr>
<th>Competencies</th>
<th>Code according to HSE Educ. Standard</th>
<th>Descriptors – main signs of mastering (indicators of achieving a result)</th>
<th>Forms and methods of teaching/learning contributing to formation and development of the competencies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Is able to explicate the scientific essence of problems in a professional field</td>
<td>СК-Б3</td>
<td>Possesses and uses</td>
<td>Lectures, practical lessons, preparation of auditorium and home works</td>
</tr>
<tr>
<td>Is able to solve the tasks in a professional field on the basis of analysis and synthesis</td>
<td>СК-Б4</td>
<td>Possesses and uses</td>
<td>Lectures, practical lessons, preparation of auditorium and home works</td>
</tr>
<tr>
<td>Is able to realize scientific and practical activities in international environment</td>
<td>СК-Б11</td>
<td>Possesses and uses</td>
<td>Lectures, practical lessons, preparation of auditorium and home works</td>
</tr>
<tr>
<td>Is able to control and develop the content of an enterprise and Internet-resources, to control the processes of creating and using informational services (content-services)</td>
<td>ПК-13</td>
<td>Possesses and uses</td>
<td>Lectures, practical lessons, preparation of auditorium and home works</td>
</tr>
<tr>
<td>Designing and using the components of IT-infrastructure of an enterprise ensuring the achievement of strategic goals and support of business-processes</td>
<td>ПК-18</td>
<td>Possesses and uses</td>
<td>Lectures, practical lessons, preparation of auditorium and home works</td>
</tr>
<tr>
<td>Consulting the clients as concerns rational choice of methods and tools for control of the IT-infrastructure of an enterprise</td>
<td>ПК-24</td>
<td>Possesses and uses</td>
<td>Lectures, practical lessons, preparation of auditorium and home works</td>
</tr>
</tbody>
</table>

4 Course Plan

<table>
<thead>
<tr>
<th>N</th>
<th>Theme</th>
<th>Total number of hours for the course</th>
<th>Auditorium hours</th>
<th>Independent work</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Lectures</td>
<td>Seminars</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Topic</td>
<td>7</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>---</td>
<td>----------------------------------------------------------------------</td>
<td>----</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>1</td>
<td>Basic concepts of the theory of applied intelligent systems and of Semantic Web</td>
<td>16</td>
<td>2</td>
<td>12</td>
</tr>
<tr>
<td>2</td>
<td>The approach of first order logic to representing information</td>
<td>30</td>
<td>8</td>
<td>14</td>
</tr>
<tr>
<td>3</td>
<td>Basic informational languages of Semantic Web project and their application to ontology design</td>
<td>33</td>
<td>5</td>
<td>22</td>
</tr>
<tr>
<td>4</td>
<td>Methods of constructing semantic representations of natural language sentences and discourses by means of the theory of K-representations</td>
<td>22</td>
<td>3</td>
<td>16</td>
</tr>
<tr>
<td>5</td>
<td>The role of linguistic databases in natural language processing</td>
<td>36</td>
<td>5</td>
<td>26</td>
</tr>
<tr>
<td>6</td>
<td>Principal approaches to semantic-syntactic analysis of natural language texts by computer systems</td>
<td>18</td>
<td>4</td>
<td>10</td>
</tr>
</tbody>
</table>

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>162</td>
<td>82</td>
<td>28</td>
</tr>
</tbody>
</table>
5 Reading List

a. Required


b. Optional


6 Grading System

<table>
<thead>
<tr>
<th>Kind of control</th>
<th>Form of control</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Current</td>
<td>Auditorium written test</td>
<td>+</td>
<td></td>
<td></td>
<td>Written test, 90 minutes, assessment of results - 1 week</td>
</tr>
<tr>
<td></td>
<td>Essay</td>
<td></td>
<td>+</td>
<td></td>
<td>Essay, the volume 12 – 15 printed pages (Times New Roman, 14 pt), assessment of results - 2 weeks</td>
</tr>
<tr>
<td>Final</td>
<td>Test</td>
<td></td>
<td>+</td>
<td></td>
<td>Written test, 90 minutes, assessment of results - 1 week</td>
</tr>
</tbody>
</table>

7 Guidelines for Knowledge Assessment

The course is studied during second and third modules. A current control is scheduled for each of these modules. Third module includes the final control – a test. The 10-balls grading scale is used for the marks in all forms of current control.

The formation of a final mark for the discipline is done as follows.

Formation of accumulated grade for 2 module

During auditorium lessons, the activity of students at lectures and practical lessons, the participation in discussions, the correctness of solutions to formulated tasks is assessed.

An accumulated grade for current control takes into account the correctness of written test (indicated in a Working Learning Plan) fulfilled by a student:
The accumulated grade (10-balls scale) for the work at practical lessons - $O_{\text{auditorium}}$.
An independent work of students is also assessed: the correctness of prepared home works (the tasks are given at practical lessons; the completeness of explicating the themes of reports.
The accumulated grade (10-balls scale) for independent work – $O_{\text{indep-work}}$.
The accumulated grade (10-balls scale) for 2 module is calculated as follows:
$$O_{\text{accumulated for 2 module}} = 0.6 \cdot O_{\text{current}} + 0.2 \cdot O_{\text{indep-work}} + 0.2 \cdot O_{\text{auditorium}}.$$

The arithmetical manner of rounding the accumulated grade is used.

*Formation of accumulated grade for 3 module*

During auditorium lessons, the activity of students at lectures and practical lessons, the participation in discussions, the correctness of solutions to formulated tasks is assessed.
An accumulated grade for current control takes into account the correctness of written test (indicated in a Working Learning Plan) fulfilled by a student:
$$O_{\text{current}} = Q_{\text{essay}}.$$

The accumulated grade (10-balls scale) for the work at practical lessons - $O_{\text{auditorium}}$.
An independent work of students is also assessed: the correctness of prepared home works (the tasks are given at practical lessons; the completeness of explicating the themes of reports.
The accumulated grade (10-balls scale) for independent work – $O_{\text{indep-work}}$.
The accumulated grade (10-balls scale) for 3 module is calculated as follows:
$$O_{\text{accumulated for 3 module}} = 0.6 \cdot O_{\text{current}} + 0.2 \cdot O_{\text{indep-work}} + 0.2 \cdot O_{\text{auditorium}}.$$

The arithmetical manner of rounding the accumulated grade is used.

*Formation of final grade for the discipline*

The final accumulated grade for the discipline is formed as follows:
$$O_{\text{final-accumulated}} = 0.5 \cdot (O_{\text{accumulated for 2 module}} + O_{\text{accumulated for 3 module}}).$$

The arithmetical manner of rounding the final accumulated grade is used.
The final grade for the discipline is calculated in accordance with the following formula:
$$O_{\text{final}} = 0.6 \cdot O_{\text{test}} + 0.4 \cdot O_{\text{final-accumulated}}$$
where $O_{\text{test}}$ – the grade for final test (a result obtained directly during a final test).

The grade for final control (a result obtained directly during a final test) $O_{\text{test}}$ is of blocking character, in case of an unsatisfactory grade it is equal to the final resulting grade for the discipline.

The arithmetical manner of rounding the final resulting grade is used.

It is possible that during the final test a student receives an additional question (an additional theoretical task), a maximal grade for an answer (or a solution) is 1 ball.
The themes of essays on the analysis of a selected direction of developing and using semantic information systems

Theme 1: The analysis of the current state of the studies on semantics-oriented computer processing of natural language.

Theme 2: The analysis of the current state of the studies on Recommender Systems with Natural Language Interface.

Theme 3: Semantic search of information in Internet and corporative computer nets.

Theme 4: The Construction of Ontologies with the Help of Natural Language Processing Systems.

Theme 5: Intelligent Text Summarization

Theme 6: Semantic annotations of electronic documents and their use

Theme 7: Semantic foundations of e-science (that is, of scientific studies based on electronic forms of representing and exchanging information).

8 Methods of Instruction

The following educational technologies are used for the realization of various kinds of teaching/learning processes: lectures, presentations of students’ reports, discussions, solutions to tasks, provision of additional possibilities of studying theoretical materials for the students wanting more deeply learn the discipline.

9 Special Equipment and Software Support

a. Software

In order to solve practical tasks and to prepare presentations, the students use modern teaching laboratory basis, including:

- standard packets of applied computer programs, in particular:
  - informational systems for the preparation of texts (Microsoft Word);
  - the systems for the preparation of presentations (Microsoft PowerPoint).

The lecturers use personal computers/notebooks and a head projector for delivering lectures and conducting practical lessons, and also use a hardware of computer classes.

b. Distance support of the course

The means of distance support provided by the system LMS are employed. The principal role in distance support is played by theoretical materials for obligatory learning, theoretical materials for optional studies, the tasks for homed works, and questionnaires,

Автор программы:

__________________________________________ В.А. Фомичев