Multi-Agent Meta-Search Engine Based on Domain Ontology

Marat Kanteev [2], Igor Minakov [1], George Rzevski [2],
Petr Skobelev [1], Simon Volman [1]

[1] The Institute for the Control of Complex Systems, RAS,
Sadovaya Str., 61, 443020, Samara, Russia.
{minakov, skobelev, volman}@magenta-technology.ru
[2] MAGENTA Technology, Gainsborough House, 59 - 60 Thames Street
Windsor, Berkshire, SL4 1TX, UK
{ kanteev, george.rzevski }@magenta-technology.com

Abstract. This article describes a new approach of HTML pages search via Internet, which is based on the semantic understanding of pages content by means of multi-agent technology. Multi-agent text understanding system, which is the basis of the approach, converts an input query and pages, received from conventional search engines, to formalized semantic descriptors, and evaluates similarity of these descriptors. What makes the approach better than the classical one and gives it an ability to cope with current search-via-Internet difficulties more effectively in the text understanding algorithm, which is capable of understanding words synonymy and ambiguity in a text, detecting and resolving sense conflicts between different parts of text, acquiring additional implicit information and working with inter-phrasal context. Semantic descriptors of the text are represented as semantic networks, which include instances of problem domain ontology concepts. Both text understanding and descriptor comparison algorithms use the knowledge about problem domain, and this problem domain is stored in the form of ontology. Ontology contains the information about the language and the problem domain of the texts used, and it can be easily updated by experts in the course of work. The approach described above was applied to the analysis of web-pages related to car industry. As a result a meta-search engine was developed, capable of analyzing pages, retrieved from traditional search engines and sorting pages by their semantic relevance to the user request. In this article one will find description of the system, testing results and system future perspectives.

Keywords: problem domain ontology, multi-agent text understanding, semantic networks comparison, web-based search, meta-search.

1 Introduction and Problem Definition

Today the World-Wide Web has become so popular that the Internet now is one of the main means of publishing information. To simplify the user access to required

information, the Internet search engines were created that allowed getting a set of Internet pages on the basis of request keywords defined by the user.

In addition to conventional information search services, the search engines recently provide the users with a new service called “advertising links”. The key point of this service is as follows: large commercial companies provide the administration of this or that search engine with links to their web-site with detailed description of information it contains. As a rule this information is presented as a set of keywords. When the user specifies in his request a keyword that belongs to this set, he gets the reference to this company’s web page. Generally this reference occupies separate space of a page with found results, it is detached from a set of resultant links. Unlike general network resources, the set of keywords for advertisement pages is created manually by the customer and search engine experts.

The main problem with using search engines is that the keywords specified by the user are generally too common and abundant. Thus the user gets thousands of useless pages but at the same time fails to find other pages with the same subject stated differently.

Therefore to make inexperienced user work efficiently with search engines, it is necessary to improve the algorithms of analyzing web-pages contents and the user queries in order to generate descriptors that contain monosemantic information on its concepts/relations. Besides comparison methods are required for these descriptors as such methods will allow evaluating similarity of two descriptors and using this evaluation to generate a set of links to the pages containing information required for the user.

2 Multi-Agent Approach – Key Points

The proposed system is a semantic meta-search engine. It means that when processing the user query it addresses this query to conventional search engines, gets search results and then analyses the contents of found pages and prioritizes them in decreasing order according to their relevance to the user query.

The main difference between existing meta-search engines is the use of problem domain ontology, which helps to represent and reconstruct semantics of request and available pages. The intelligent multi-agent algorithm of analyzing texts in the natural language is used for this purpose.

Multi-agent approach is developing rapidly in the recent time in a number of applications such as distributed systems, logistic and supply chain scheduling, data mining etc., as described in [1]. New line of multi-agent systems investigation, which is studying open non-equilibrium systems with energy income, as described in [2], is also very important. It was shown in [3], that solution methods of the resource distribution task in open systems for logistic applications can be effectively used as well in the knowledge management, data clustering, natural language processing systems etc. For the text understanding systems, open non-equilibrium state means that the sense of previous words and phrases can be changed and revised during next phrases analysis.

The essence of the used algorithm is that each word of the text is assigned an autonomous software agent capable of negotiating with other similar agents about the meaning of each word in the sentence and its general meaning on the basis of domain ontology.

During negotiations the word agents can speculate about the possible word meanings and their semantic relations, find and resolve meaning conflicts, detect implicit information on the basis of domain knowledge, take into account the context of the word usage within one sentence and inter-phrasal context thus connecting the words of various sentences into one semantic network, as described in [4].

During negotiations the agents use the knowledge about the problem domain in use and language contained in ontology. Ontology is the variety of semantic networks that stores information about the concepts, attributes and relations used in the problem domain. Ontology can be edited by adding new rules and concepts on the fly. Definition of ontology can be found in [5].

For example in case of tourism ontology the following sequence of phrases is being analyzed: ”I want to travel to London. My wife and my son will go with me. I would like to book a three-room apartment with the sea view”. In the course of the phrase analysis the scene of the world will be built - here ”I” will be recognized as ”man”, ”London” – as a ”city”, and these two concepts will be related by ”going to travel”. When second phrase comes into the system, –the concepts of two more persons, such as the ”wife” and the ”son” appear on the scene. Here ”I” will be related to ”wife” by ”husband-wife” type of relation and to ”son” – by ”farther-son” type of relation. Moreover, ”wife” and ”son” will also be related to ”London” by ”going to travel” type of relation. The coming of the third phrase will clarify the scene even more, as the apartment is more likely to be in a hotel, and the hotel – in London.

As the result of analysis of text in the natural language we get text descriptor, which contains formal monosemantic description of the initial text meaning. Descriptor represents a semantic network that consists of ontology concepts and their relations. Unlike the initial text represented as a line it is rather simple to compare such descriptors on similarity of information they contain on the base of the ontology.

The algorithm of comparing semantic descriptors was developed for multi-agent meta-search engine. This algorithm is based on finding in one of the descriptors the sub-network which is close to the network of the other descriptor as much as possible. Similarity degree of two sub-networks is defined as similarity degree of their respective pairs of nodes. Similarity degree of two nodes depends on relative position of corresponding concepts in the ontology and on the values of attributes connected with nodes under comparison. Speed of this algorithm is about 2000 descriptors per second.

Application of the algorithms of building and comparing the formalized meaning of text in the natural language allows eliminating the key problems that appear when searching web-pages in the Internet:

Thus the system we offer allows solving such problems, typical for traditional systems, as:

- **Incomplete information in text of the pages**. The algorithm of page processing and generating semantic descriptors allows extracting additional implicit information on the basis of domain knowledge stored in the ontology. The system

also provides possibility of manual creating and editing page descriptors thus improving their reliability.

- **Synonymy and homonymy of language symbols and dependency of the word meaning on the context of its usage.** Descriptors are monosemantic structures consisting of instances of domain concepts (denotations). The problem of ambiguous relations between words and concepts is solved at the stage of text analysis using the word context and knowledge about the used problem domain.
- **Finding semantic dependency between various sentences and page sections.** The algorithm of creating descriptors can process correctly inter-phrasal relations of words using domain ontology. As the result descriptor represents a connected network that unites concepts mentioned in various parts of a page.
- **Cutting off non-significant words of a query.** Semantic descriptors of a query receive only the most important concepts from problem domain (i.e. those contained in the ontology) and do not receive not informative common words

3 Multi-agent text understanding - the proposed method

The method consists of the following four steps (see Figure 1).

- Morphological analysis
- Syntactic analysis
- Semantic analysis
- Pragmatics

The process goes as following - all text is divided into sentences. Sentences are fed into the meaning extraction process one by one. And for each sentence first three stages are applied. When all text is parsed, then resulting semantic descriptor goes into forth stage – pragmatics.

Morphological Syntactical Analysis

An agent is assigned to each word in the sentence;

- Word Agents access Ontology and acquire relevant knowledge on morphology
- Word Agents execute morphological analysis of the sentence and establish characteristics of each word, such as gender, number, case, time, etc.
- If morphological analysis results in polysemy, i.e., a situation in which some words could play several roles in a sentence (a noun or adjective or verb), several agents are assigned to the same word each representing one of its possible roles – i.e. creating several branches of possible sentence parsing results.

Syntactical Analysis

- Word Agents access Ontology and acquire relevant knowledge on syntax;
- Word Agents execute syntactical analysis where they aim at identifying the syntactical structure of the sentence. For example, a Subject searches for a Predicate of the same gender and number, and a Predicate looks for a suitable Subject and Objects. Conflicts are resolved through a process of negotiation. A grammatically correct sentence is represented by means of a Syntactic Descriptor;

• If results of the syntactical analysis are ambiguous, i.e., several variants of the syntactic structure of the sentence under consideration are feasible, each feasible variant is represented by a different Syntactic Descriptor. If none syntactic descriptor is valid, then several with high enough level of correctness (i.e. more than 80% of grammatically possible links present in the syntactic descriptor) are selected for the next stage.

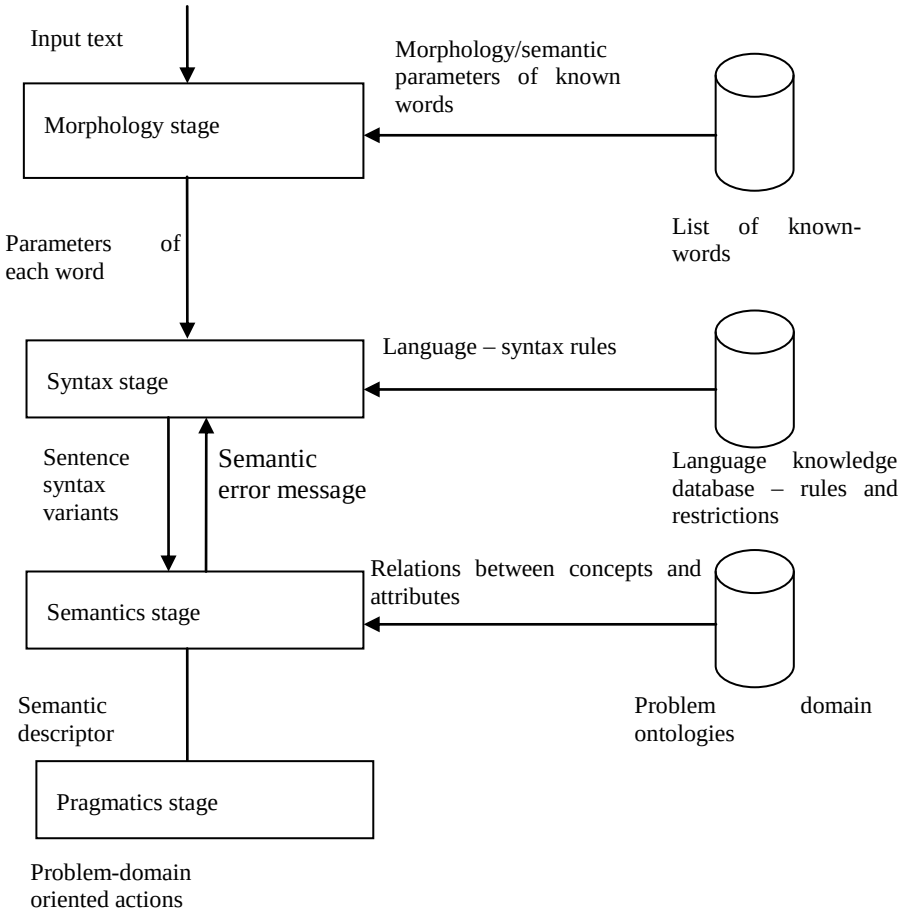

Fig. 1. General scheme of NLP

Semantic Analysis

• Word Agents access Ontology and acquire relevant knowledge on semantics – including possible relations between concepts and valid attribute values;

• Each selected grammatical structure of the sentence under consideration is subjected to semantic analysis. This analysis is aimed at establishing the semantic compatibility of words in each grammatically correct sentence. Word Agents learn

from Ontology possible meanings of words that they represent and by consulting each other attempt to eliminate inappropriate alternatives by building at least contradictive semantic descriptor basing on problem domain ontology.

- Once agents agree on a grammatically and semantically correct sentence, they create a Semantic Descriptor of the sentence, which is a network of concepts and attribute values contained in the sentence;
- If a solution that satisfies all agents cannot be found, agents interact with the user explaining the difficulties and suggesting how the issues could be resolved or select most probable decision automatically, if level of correctness is high enough (i.e. more than 80% links are valid according to the problem domain ontology);
- Each new grammatically and semantically correct sentence generated by the previous steps is checked for semantic compatibility with Semantic Descriptors of preceding sentences. In the process agents may decide to modify previously agreed semantic interpretations of words or sentences by returning to earlier stages of negotiation (described above) with its new knowledge, which improve confidence in certain option and will result in reconstruction of Semantic Descriptor for preceding sentences;
- When all sentences are processed, the final Semantic Descriptor of the whole document is constructed thus providing a computer readable semantic interpretation of the text

Pragmatics

- Word Agents access Ontology and acquire relevant knowledge on pragmatics, which is closely related to the application at hand;
- At this stage agents consider their application-oriented tasks and decide if they need to execute any additional processes. For example, if the application is a Person – Computer Dialog, agents may decide that they need to ask the user to supply some additional information; if the application is a Search Engine, agents will compare the Semantic Descriptor of the search request with Semantic Descriptors of available search results. If the application is a Classifier, agents will compare Semantic Descriptors of different documents and form groups of documents with semantic proximity.

Let us recapitulate main features of the proposed method.

- Decision making rules are specified in ontology, which incorporates general knowledge on text understanding, language-oriented rules and specific knowledge on the problem domain;
- Every word in the text under consideration is given the opportunity to autonomously and pro-actively search for its own meaning using knowledge available in ontology;
- Tentative decisions are reached through a process of consultation and negotiation among all words;
- The final decision on the meaning of every word is reached through a consensus among all words;
- Semantic Descriptors are produced for individual sentences and for the whole text, thus representing semantics of the coherent document;
- The extraction of meanings follows an autonomous trial-and-error pattern (self-organization);
- The process of meaning extraction can be regulated by modifying ontology.

4 Meta-Search Engine Architecture

The general system structure is shown in Figure 2.

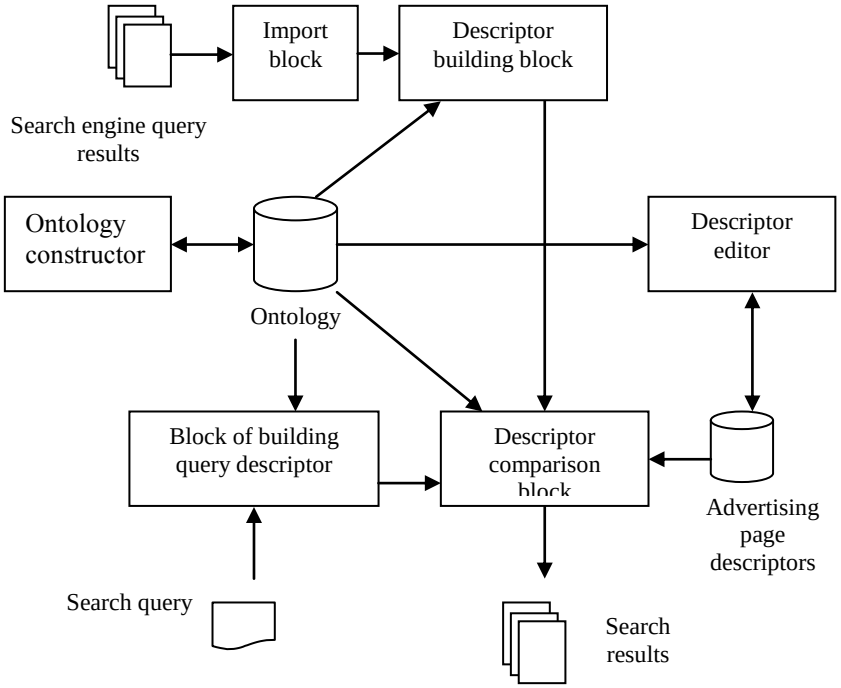

Fig. 2. Meta-search Engine General Structure

The system contains the following main components:

- **Import Block:** provides import of data into the system and transforms the data into the text.
- **Blocks of building page and query semantic descriptors:** automatically convert text information to semantic descriptor on the basis of domain-specific knowledge.
- **Block of comparing semantic descriptors:** compares two semantic descriptors built on the basis of domain-specific knowledge.
- **Ontology:** stores the system knowledge used when processing a query.
- **Ontology constructor:** provides possibilities of editing the system ontology.
- **Descriptors of advertising pages:** stores semantic descriptors of advertising pages.
- **Descriptor editor:** provides possibility of editing semantic descriptors on the basis of ontology.

In general the system workflow can be divided into 2 stages. At the first stage an expert user creates a domain ontology (or can use the existing one). This ontology serves the basis for automatic text analysis and contains semantic relations, syntactic

rules and morphological tables of words. The ontology is created in semi-automated way, where system is provided by a set of documents from a selected problem domain, and all words, excluding non-meaningful, are presented to the user. Furthermore, for each word, basing on its frequency in texts and relative positions comparing to other words initial guess is made, whether it's an object, relation or attribute value. After that user is able to manually assign each word as an object, relation, attribute or attribute value, and link concepts between themselves. Further on the ontology can be continuously extended with new common words and terms during the process of work.

At the second stage, when the user search query arrives, this query is redirected to conventional search engine. At the same time the module of text processing starts building semantic query descriptor. Currently a creation of semantic descriptor for average site takes approximately 2-3 seconds, but for the set of documents this process can be easily made parallel, as process is independent for each document.

After getting the resultant set of pages from a search engine the definite number of pages (10, 50, 100, etc.) that the search engine considers to match the query best of all is selected. Then the developed meta-search system builds semantic descriptors of these pages and compares them with semantic descriptors of query. As the result the pages are rearranged. According to comparison results, the pages that fit the query most of all, are placed on the top of the list.

Besides the query descriptor is compared to descriptors of advertising pages, stored in the system database. The appropriate pages are also shown to the user in a separate section of results window.

5 Example of Search Query Processing

Let us specify the system workflow. As has been mentioned above, the first step is building domain ontology. E.g., the ontology containing the concepts "car", "manufacturer", "car body", "engine", etc. was created for the system under consideration (cars domain). This ontology included 60 objects, 40 attributes, 10 classes of relations and near 2000 instances of relations (including inherited ones). See figure 3.

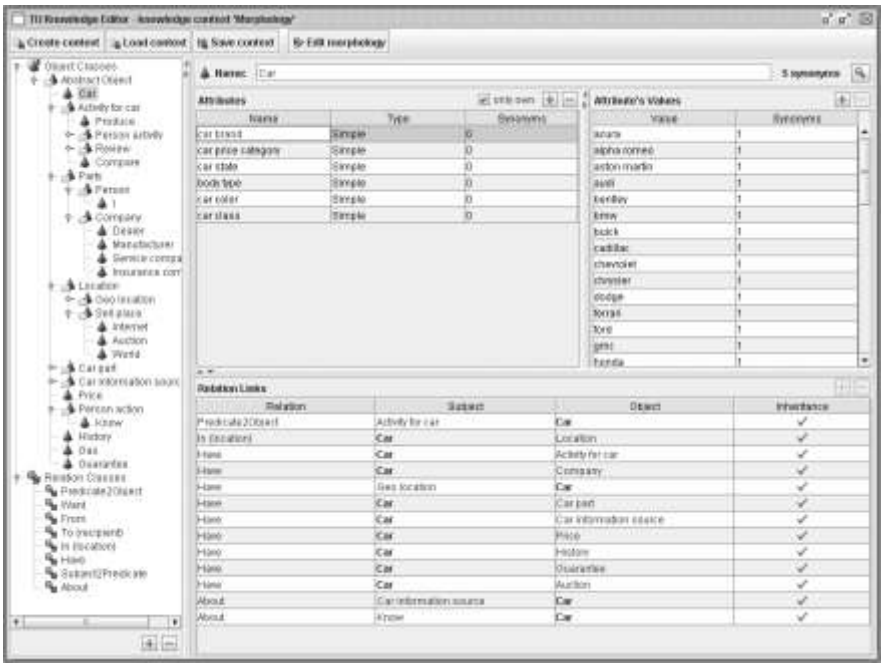

Fig. 3. Example of Ontology

Then descriptors of advertising pages were built. It was made semi-automatically: first of all the page content was analyzed by the block of automatic descriptor building, if necessary descriptor was edited, in order to improve the accuracy of stored information about the page.

For example, a user has sent the query "Cars magazines in US". The block of query descriptor building created the following query descriptor – see Figure 4.

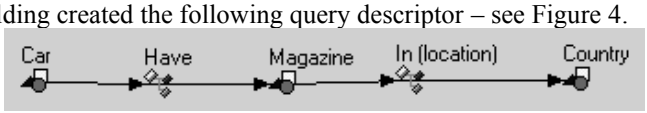

Fig. 4. Example of descriptor

After that the query was directed to Google search engine that returned near 64 mln links (09 October 2006) – see Figure 5.

Fig. 5. Google Query Results

The first ten of them, given in the order they were returned by Google, are as follows:

1. Car and Driver Magazine : Homepage (http://www.caranddriver.com/);
2. Automobile Magazine (http://www.automobilemag.com/);
3. Motor Trend - Car, Truck, SUV Road Tests, Buyer's Guide, News (http://www.motortrend.com/);
4. Popular Mechanics PM Zone--Automotive, Technology, Home Journal .(http://www.popularmechanics.com/);
5. BMW Car Magazine (http://www.bmwcarmagazine.com/);
6. Portal: Car Magazines & Car Publications (www.car-stuff.com/carlinks/pub.htm);
7. TraderOnline.com Classified Ads (http://www.traderonline.com/);
8. Autoweek.com The online version of the weekly auto magazine (http://www.autoweek.com/);
9. Online shop of model cars, videos and magazines(www.ewa1.com/);
10. Import Tuner Magazine (http://www.importtuner.com/).

Having checked the returned links, we can say that most of them refer either to some car magazines or to mass media in general (US News), or to some pages that cover the topic “cars” but are not connected with magazines at all. The only relevant links

(Car Magazines and Automotive News, Auto Racing and Car Magazines Directory) are two last ones.
After comparing descriptors of found pages with query descriptor, the system gave the following rating of relevance (see Figure 6).

Title	URL	Rating
DIECAST MODEL CARS, 1/43, 1/18, EWA, CA...	http://www.ewa1.com/	1020
CAR MAGAZINES & CAR PUBLICATIONS (CAR...	http://www.car-stuff.com/...	820
NEW CARS, CAR REVIEWS & PRICES, USED ...	http://www.motortrend.com/	610
BMW CAR MAGAZINE FROM UNITY MEDIA	http://www.bmwcarmagazi...	610
TUNER CARS, IMPORT MODELS AND PICTU...	http://www.importtuner.com/	610
NEW CARS, CAR REVIEWS, CONCEPT CARS,...	http://www.automobilemag...	610
POPULAR MECHANICS	http://www.popularmecha...	410
CAR AND DRIVER - 2006 CAR REVIEWS AND...	http://www.caranddriver.c...	400
AUTOWEEK	http://www.autoweek.com/	400
TRADERONLINE.COM CLASSIFIED ADS	http://www.traderonline.co...	300

Fig. 6. Query Results

We can see that the system placed the most relevant links on top of the list and their rating is more or less the same exceeding the rating of other pages nearly 2 times.

6 Analysis Method and Experiments Results

A number of test experiments were made to analyze the quality of this program.

We've taken most popular examples of requests in car industry domain, according to Google and Overture statistics. In average we analyzed approximately 100 typical requests, 100 pages per request per search engine. Below we are giving figures for most common and simple examples.

A list of 6 popular test queries is:

1. Car purchasing
2. Car rent
3. Car repairs
4. Audio systems
5. Car pictures
6. Car overview

Then each query was sent to Yahoo! search engine and the first 100 results were analyzed using multi-agent meta-search engine. After manual analysis of results the expert put into the table information concerning how many relevant documents contained the first 50 returned by traditional search engine and those returned by meta-search engine. The results are given in the table below:

Table 1. Comparing Effectiveness of Search Using Keywords and Semantic Descriptors

Query	Search using keywords	Search using semantic descriptors
Query #1	64%	74%
Query #2	70%	90%
Query #3	62%	84%
Query #4	60%	88%
Query #5	44%	68%
Query #6	52%	72%
Total	58.67%	79.33%

Next, advertising links returned by Google and meta-search engine for the same queries were also analyzed in terms of their relevance to the query. Each reference was considered by the expert to be either relevant or not relevant. Results are given in the table below:

Table 2. Comparing Search Modes Using Advertising Pages

Query	Search using keywords		Search using semantic descriptors	
	Quantity	Relevance	Quantity	Relevance
Query #1	8	75%	15	87%
Query #2	8	75%	4	100%
Query #3	2	50%	3	100%
Query #4	5	100%	3	100%
Query #5	2	100%	2	100%
Query #6	0		7	86%
Total	4.17	80.00%	5.67	95.50%

To summarize all our analysis, meta-search system proved to be more effective than conventional search engines and traditional meta-search systems. During automatic generation of information about the page (usual search mode) the average number of relevant pages was 20-22 % higher than that of traditional search engines. During manual generation of information about the page (advertising links) the difference was a little bit lower – 15-16%, because of more thorough selection of keywords for advertising pages in traditional search engines.

7 Conclusion

Therefore, the following results were achieved:

- Methods of editing, storing and comparing formalized descriptors of the text meaning were developed providing possibilities of eliminating main problems of working with texts in the natural language;
- The algorithm of generating these descriptors on the basis of text in the natural language was developed providing automatic construction of web-page semantic descriptors and descriptors of user input queries;
- Initial draft version of meta-search engine was developed providing interaction with the most popular search engines, storage and editing of its own set of advertising pages. It also provides debugging interface that allows to control the process of generating and comparing semantic descriptors and get information about decision making of the system;
- Application of this system allows improving the effectiveness of searching text information in the network and other sources, increasing relevance of results, especially for those who doesn't have enough experience of using advanced methods of search engines.

References

1. Wooldridge M.: An Introduction to MultiAgent Systems: New-York, Jonh Wiley & Sons (2002)
2. Nicolis G., Prigogine I.: Self-Organization in Nonequilibrium Systems: New-York, Jonh Wiley & Sons (1977)
3. Skobelev P.: The Theoretical Basis of the Open MAS Development for Operational Data Processing in the Decision Making Process. In 5th International Conference 'Complex Systems: Control and Modelling Problems', Samara, Russia (2003) 295 – 303
4. Andreev V., Iwkushkin K., Karyagin D., Minakov I., Rzevski G., Skobelev, P., Tomin M.: Development of the Multi-Agent System for Text Understanding. In 3rd International Conference 'Complex Systems: Control and Modelling Problems'. Samara, Russia (2001) 489 – 495
5. Andreev V., Iwkushkin K., Minakov I., Rzevski G., Skobelev P.: The Constructor of Ontologies for Multi-Agent Systems. In 3rd International Conference 'Complex Systems: Control and Modelling Problems', Samara, Russia (2001) 480 – 488.

Published: Kanteev M., Minakov I., Rzevski G., Skobelev P. Multi-Agent Meta-Search Engine Based on Domain Ontology. - International Workshop "Autonomous Intelligent Systems: Agent and Data Mining" (AIS-ADM 2007) - St. Petersburg, Russia, June 3-5, 2007. Volume 4476 LNAI, 2007, Pages 269-274.