ALGORITHMIZATION OF SEARCH OPERATIONS IN MULTIAGENT INFORMATION-ANALYTICAL SYSTEMS

ANASTASIA GENNADIEVNA ANANIEVA
National Research Nuclear University "MEPhI"
Kashirskoe Shosse, 31, Moscow, 115409, Russia
ALEXEY ANATOLIEVICH ARTAMONOV
National Research Nuclear University "MEPhI"
Kashirskoe Shosse, 31, Moscow, 115409, Russia
ILYA URIEVICH GALIN
National Research Nuclear University "MEPhI"
Kashirskoe Shosse, 31, Moscow, 115409, Russia
EVHENIY SERGEYEVICH TRETYAKOV
National Research Nuclear University "MEPhI"
Kashirskoe Shosse, 31, Moscow, 115409, Russia
DMITRY OLEGOVICH KSHNYAKOV
National Research Nuclear University "MEPhI"
Kashirskoe Shosse, 31, Moscow, 115409, Russia

ABSTRACT

The exponential growth in the number of data sources and data on the Internet leads to necessary selection of optimal searching engines for data collection that corresponds to user's request. The task to create searching algorithms lies primarily in the decomposition of data types, selection of appropriate information sources and adjustment of searching engines (usage of specific syntax) for collection of large amounts of data from reliable sources. Since 2008 at the department "Analysis of competitive systems" NRNU MEPhI are held activities on setting up specific agent-based searching systems and creation of unique searching algorithms and techniques for analyzing the output technological developments related to advances in the area of scientific and technological sciences. In particular, the resource cluster was collected and thesauruses were compiled on such advanced scientific fields as "Plasma physics", "Nanotechnology" and "Laser technology".This article describes the algorithms, approaches and methods of data search retrieval from the Internet that are developing and testing at the department №65 "Analysis of competitive systems" NRNU MEPhI under the "multi-agent information and analytical systems in the area of scientific and technology science" MIAS [1].

Keywords: *Multi-agent system, Big Data, Data mining, retrieval system, information and analytical system*

1. INTRODUCTION

The emergence of network communication technologies has significantly affected the speed rate and volume of various data transmission around the world. The emergence specialized local networks and as a consequences – the World Wide Web have defined the transition of data usage from print sources to electronic sources.

The XXI century marked a transition to new era of information technology, the essence of which is the emergence of data sources and sharing usage of them. New ways of storing and sharing information have become cloud storage or clouds, as well as the Internet of Things (IoT) [2][3]. The great importance for development and growth of the information and communication technologies (ICT) industry could lead to further transition of so called "Third platform", the key components of which are mobile computing, cloud services, analysis of large data sets and social networks[4].

According to forecasts of one of the largest analytical companies in the information technology market – "International Data Corporation" (IDC), - in 2015, the investment in the field of information and communication technologies all over the world will grow by 3.8% (3.8 trillion US dollars). The industry growth is entirely almost provided by the technologies of "Third platform" [5].

In such manner, significant changes are likely to happen in the following areas: business intelligence and processing large amounts of data. Global disbursements on software, hardware and services that allow to receive and process large amounts of data from a network source rapidly, will reach $125 billion. Multimedia analytics (video, audio and images) will become an important factor in stimulating the growth of projects in the area of large amounts of data processing [6]. Data as a service will become increasingly important as providers of cloud platforms and analysts will offer to their customers a qualitatively new information obtained from the public resources and limited access resources [7].

As can be seen from the above, the tasks of searching, collecting and analysis information from large amounts of data and tasks of data transformation to available analytical reports for the end user are extremely relevant for today [8][9].

Since 2008, at the department of “Analysis of competitive systems” NRNU MEPhI are held activities on the development of new generation systems, namely multi-agent systems that allow to receive data from information sources on the Internet by automated mode and carry out its primary analysis. In 2014, by the status of experimental-industrial exploitation was launched “Multi-agent information and analytical systems in the area of scientific and technology science” (MIAS), currently serving users on such thematic fields: “Plasma physics”, “Laser technologies” and others.

As part of the work was Certificate of state registration of database #2014620346 “International research and technology organizations in plasma physics” [10].

The research results were also presented at the Public Chamber of the Russian Federation on the 32nd meeting of All-Russia scientific seminary on nanotechnology organized by Nanotechnology society of Russia (NSR) and NRNU MEPhI in November 2014 and presented at scientific schools “GRID and Advanced Information Systems” at the Joint Institute for Nuclear research, Dubna.

2. DETERMINING THE TYPES OF SEARCH OBJECTS AND INFORMATIONAL FIELD RESEARCH

The base of the searching relevant data process is a determination of the essences of the search object and identification of its main attributes for inclusion in the database of MIAS, in accordance with the database structure. The task of searching and collecting relevant data itself is not a trivial one, because it requires a compilation of the unique dictionary of keywords on the specific thematic fields, in other words compilation of thematic thesaurus, for searching set of informational sources [11]. For full coverage of the information field, thematic thesaurus must be compiled in several languages in accordance with the languages of states in which information area searching is held. The set of informational sources that was previously received are used for setting up the search engine spiders (crawlers) [12].

The main advantage of the MIAS lies in the flexibility of crawlers’ adjustment. Crawlers can be of three types: crawlers that use RSS for receiving information, customizable crawlers on the principle of identifying the invariable part of the news section by pointing the patterns of alterations in the address bar and crawlers that can be configured for complex structured websites with the designation of news blocks, this option includes detailed information source html code disassembly using the software module of "regular expressions" as defined by the structure and patterns of text.

The second advantage is the organization structure of the database of the collected data in accordance with the tasks. At the core of the database (DB) is a standard relational modeling approach - entity-relationship. Entities in the database can act as the standard objects (persons, organizations, projects) and non-standard – such as events. In this regard specialists set up crawlers and rubricating process in accordance with all the objects that are stored in the database [13].

As a result of search, research and scientific activities the following key entities were revealed:

1.Technology (Scientific field);

2.Project;

3.Organization;

4.Person.

During MIAS adjustment for every particular case the boundaries of the information field are determining in accordance with the documents’ types that could be relevant to user’s request and could contain the largest amount of significant information. In practice this is official documentation and information, uploaded to reliable information sources. Under reliable resources the authors mean information resources of organizations that conduct research and development in the thematic field requested by the user.

In such manner, for the entity “Person” significant data are: information about the places of study and work, participation in projects, programs and competitions (especially provided by the government), published patents, articles or monographs, research and lists of colleagues within the same department of this person.

For the entity “Organization” significant data are: information about the official name and working area, registration place and date, administrative structure, budget, completed projects – both commercial and research, the share of research in total industrial activity, production, the number of completes state orders at the time of data collection.

For the entity “Project” significant data are: name, deadlines, funding, funding agencies, customers, general contractors and subcontractors, the main stages and the work performed, exploratory work and the main participants (project managers).

For the entity “Technology” significant data are: the essence of technology, innovation, the problem-solving technic, implementation project companies’ reports related to the usage and implementation of the technology.

The information sources of such types of entities can be official sites of the companies that publishing news about their projects and achievements, state patent systems, abstract databases (such as Web of Science and Scopus), state systems on providing tendering and competition, as well as systems in which details of all projects funded by various departments are published, specialized communities and the sites of major news outlets and social networks [14].

It is important to note that in each case the specialist should know not only the appearance and composition of the published data, but also know the keywords and codifiers for specific data on the resource.

3. METHODOLOGY OF THE SEARCH OPERATIONS AND EVALUATION OF THE QUALITY OF INFORMATION

3.1.Compilation of multi-language thematic dictionary of terms under tasks in hand

It is necessary to define the scope of the thematic search, and make a list of keywords.

This is a preparatory stage for the search of information resources and the primary information noise filtering.

Under the research thematic thesauruses were compiled on thematic areas such as “Plasma Physics” (more than 250 terms and definitions in Russian, English and Chinese), “Laser Technology” (more than 90 terms in Russian, English and Spanish) “Nanotechnology” (more than 100 terms in Russian and English), and “Dual use technology” (more than 200 terms in Russian and English).

3.2. Selection of Information Retrieval System (IRS)

Information Retrieval System (IRS) selecting is based on two main parameters: the power of data indexing on the Internet and the absence of mechanisms to collect data about the user.

Often, a person who does not have special knowledge in solving problems related to the search for information on the Internet use the most common IRS such as “Google”, “Yandex”, etc. However, there are a number of factors that must be considered [15].

Firstly, ranking sites mechanisms of IRS may comprise a different number of parameters. For example, “Google” use such parameters as:

1. Domain (domain age, history of the domain, keywords, placement of domain name, IP domain, placing the IP address / server history of sanctions domain, domain registration with the help of «Google local», the accuracy of registration data domain);
2. Server (sites inaccessible, sites with long response);
3. Information about the owner of the domain;
4. The architecture of the site (including the use of encryption protocol such as SSL, TSL and etc., and “Google Search Appliance”);
5. Content (keywords, semantic information, taxonomic flags, keywords, data of social networks «Twitter», «LinkedIn», «Facebook», etc.));
6. Inbound links (including the relevance of page content, country domain, the importance of the domain (.edu, .gov), the location of the server);
7. Outgoing links (including a keyword in an external link from the site);
8. The cluster of references (including keyword in internal reference);
9. Internal links;
10. The sanctions (including over-optimization, purchased links, corrupt links, spam);
11. Searching factors (including bounce rate, search for a domain name, or brand, users from which countries come to the site);
12. Metadata and site performance.

It is important to consider the fact that each IRS has its own ranking algorithms for searching results. Under this statement authors mean that the policy of such large companies as “Google” and “Yandex” is aimed to collecting data about the user, this information is used for ranking algorithms that leads to different outgoing searching results in case of usage the same queries from different IP addresses and this leads to subjective searching activities. In order to overcome this problem it is suggested to apply to several IRS including such systems as “Duckduckgo”, as the policy of these IRS is aimed to giving the same ranking algorithm without collecting data about the users [16].

While searching information in foreign language it is suggested to apply to national IRS of the state which language is used. For instance, while searching data in information field of China in addition of using “Google”, “Yandex”, “Duckduckgo” and etc., it is suggested to refer to the domestic IRS of China – “Baidu” [17].

It is clear that for the most relevant information cluster to task in hand it is necessary to estimate the specifics of various types of IRS, as well as to use different settings to access the possibilities of IRS, particularly techniques of IP-addresses concealment and usage of browsers that can help to reduce the possibility of defining any personal user information [18].

3.3.Compilation of search queries using advanced search syntax

Technologies include the use of an advanced search query language elements, such as quotation marks "" operators «OR», «-», «...», as well as the choice of the domain, file format, country, date, language, rights to use and safe search. Using the advanced search syntax with the knowledge of certain codifiers assigned to various types of official documents, greatly facilitates the process of searching relevant data on the first phase of the cluster set of information sources.

It is important to note that the usage of Google domain (web site) search gives more results that native search engines of these sites which is determined by large corporations metadata capabilities of the search robots, which leads to greater coverage of information sources on the Internet [19].

3.4. Organizations’ web-sites selection that publish request relevant information

The important stage of selection of monitored source for adjustment process of crawlers in the MIAS is to evaluate the reliability of the source itself. At the moment, there are two criteria for evaluating sources:

- The resource is registered to a trusted organization or person;
- The resource has the highest rating values in different systems, for example, Alexa.com.

The combination of these two criteria minimizes the possibility of false set of information sources in the first stage. However, in the case of controversial issues it is required additional research on the reliability of the resource parameters such as: the date of registration of the resource, regular updates, no non-targeted advertising on the site and the list of partner sites, which also has links to the resource [20].

3.5. Adding web-site of the organization proposed by the user to the search cluster of information resources (optional step)

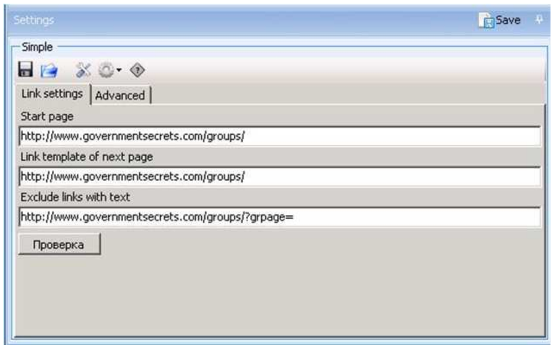
The implementation of this phase is possible when there is an interactive mode with the information user. This step is important as it helps to minimize the efforts required for search and analysis of information resources. However, in the course of the thematic areas “Plasma physics” and “Laser technology” has been revealed that the obtained cluster from user is minimal and cannot cover all the specific directions on one thematic sector.

3.6. Crawlers adjustments on informational resources in accordance with the schedule given by the expert

After selecting relevant sources of information, expert perform an adjustment of crawlers on collecting data and adjustment for further data rubricating process. This process is described in detail in Section 3.

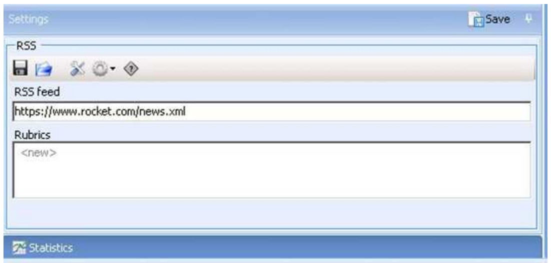
Here is an example: crawlers were set up on collecting data from the compiled set of monitoring information source cluster on the subject of “selection of foreign news in the field of dual-use technology” by the expert. The number of monitored information source in the MIAS are unlimited.

The examples of crawler adjustment for collecting data from Aerojet.com and Governmentsecrects.com are shown on image 1 and image 2.



Start page: http://www.governmentsecrets.com/groups/
Link template of next page: http://www.governmentsecrets.com/groups/
Exclude links with text: http://www.governmentsecrets.com/groups/?grpage=

Fig 1.RSS Adjustment In MIAS



RSS feed: https://www.rocket.com/news.xml
Rubrics: <new>

Fig 1. Adjustment Of Simple Template

3.7. Data searching on unique codifiers and documents types

Currently, there is many different information codification systems (ISBN, ISSN, DOI, DUNS, NAICS, etc.). The techniques of their generation are based on sequential coding object specifications and classification that are subject to encoding (subject area of research and publications, the State or a language of publication, the area / field of activity, as well as random numbers). The usage of codification systems allows to significantly improve search effectiveness and speed rate because of the unique nature of some codes that leads to low probability of getting non relevant information on the outcome compared to the searching using keywords.

After performing the steps above the researcher receives first cluster of monitored sources and adjusted crawlers in accordance with searching requirements. The researcher receives the cluster of information from the MIAS database and starts analyze it.

4. DATA SYSTEMATIZATION AND RUBRICATING PROCESS

As soon as first information cluster have been obtained from relevant and reliable sources of information it is possible to start analyzing this information cluster. This process can be divided into several parts:

4.1. Information rubricating process;

4.2. Compilation of relation map that include relations between research objects for initiating deep search;

4.3. Downloading information from the MIAS database that was brought by crawlers;

4.4. Analysis, filtering from information noise and data structuration in accordance with task in hand;

4.5. Data systematization on preparing reports;

Data rubricating process performs in accordance with search entities types, for example input data about organization, person, project and aggregation of all news blocks on thematic area in hand. The MIAS machine algorithms allow to identify relations between two different entities and keywords in the information clusters by autonomous mode. That allows expert perform additional analysis on outgoing data and develop semantic nets, that are presented as graphical reflection of the objects' specifications and relations between them.

After holding analysis there is a sense to repeat steps 3.1 – 3.7, that were described in Section 3 for updating cluster of relevant information sources.

The data systematization for end report preparation means good imagining of text and graphical data of received information and relations between object in a comfortable manner for the end user. In this moment the department experts have developed procedures for preparing further report types:

- Digests on news and events;
- Object dynamic profile.

The digests on news and events are presented as regular issue of relevant to particular request information with a different rubricating structure, for instance rubricating based on differentiation of states and/or technologies and thematic fields. In news digest more preferable data format are articles with graphical objects (images, photos, schemes) that were received from the monitored information source. The digest contain statistics of keywords that were indicated by the user and the list of information sources.

The object dynamic process is more interesting document from the researchers' point of view as it included such sections as: data about research object, dynamics of development of the object since emerging to the moment of data collection, financial information, relations to other objects, researches and patents, statistics and perspectives of object development. The data visualization could be presents as a table or semantic web that are prepared individually to every request.

In this activity the stages of project VTOL X-PLANE were researched with the beginning of pre-research analyze to achievements analyze of the reporting year of the project. Under the research there were revealed relations between participants and performers of the project, budget, technology development and main project results. The figure presents a schematic summary of the results of the primary object of study - a person, organization, technology, and communication between them.

5. RESULTS AND DISCUSSION

Existing traditional models for search allow identify information sources by thematic field with the help of keywords. This work requires significant time-consuming because of noise information filtering process.

Let us have a list of N sites, which should be monitored for the search of new information. The time required for a full cycle analysis of sources in one session of monitoring is given by:

$T_{gen} = N(t_{site}+M*t_{evaluation})$, where (1)

T_{gen} – time spent on analysis of sites on one thematic field in one session of monitoring;

N – the number of sites of thematic field;

T_{site} – time spent on processing one site;

M – the number of new news massages on one site;

$T_{evaluation}$ – time spent on identification relevant information of news massage on one site.

Equation (1) determines the time-consuming for one session of monitoring by traditional means.

By the equation, the time for information searching and relevance evaluation is the amount of time – the time spent on visiting site and time spent on evaluation of relevance of a massage. Thus the assessment of the costs T_{gen} time increases linearly with the number of sources of thematic information and the number of news reports.

Experimental studies have shown that the various thematic areas of the value of the size of the cluster sites with professionally significant information is several tens of, in this case 60.

The number of news massages between two sessions of monitoring comprises an average 50 units.

These values are random, so the first approximation can be assessed by mathematical expectation or universal mean.

By the experiment result, T_{gen} has further values:

N = 60 – the number of site from thematic field;

t_{site} =2 min. – time spent on processing one site;

M = 50 – the number of news massages for the session of monitoring;

$t_{evaluation}$ = 8 min. – time spent on evaluation of relevance of a new massage on a site.

Therefore, in view of the experimental data, the average time for one monitoring session is about 24120 minutes or 402 hours, amounting to 50.25 man-days at 8 hour day per subject area.

At such time expenses a specific person has to be engaged in information retrieval or the expert would not be informed enough on its thematic field.

Performing the same operation by agent system, with preconfigured crawlers on the thematic field take 15 minutes – 0.25 hours.

The comparison shows that process is optimized by three orders of value. In this case the user is fully informed with spending time only on analyze of the messages that are interesting to him without time spending on the data collection.

The digests on the thematic fields “Plasma physics” and “Laser technology” are completed with the usage of MIAS instruments for 24 man-hours. In this case digest release is limited by objective temps of information updates on a thematic field but not by formation and release.

The task on collecting, storing and analysis of large amount of data is currently central. However, even with the existing variety of IRS and special programmable instruments for collecting data, there is a difficulty in selecting techniques of data collecting relevant to users’ request and data evaluation.

6. CONCLUSION

This article describes examples of works in which some techniques were used that allow the target setting of the multi-agent systems on searching data on the Internet as well as creation of object-oriented databases in the MIAS by the type researched objects and perform regular collection and analysis of large amount of data.

It is planned to enhance the MIAS by optimizing searching operations of data collecting and formation of keywords databases and research objects – person, organization, thematic field and technologies.

REFERENCES:

[1] Budzko V., Leonov D., Nikolayev V., Onykiy B. and Sokolina K. (2011). Development of Information and Analytical Support for Scientific and Research Activities in the National Nuclear Research University «MEPHI», Highly available systems. 2011. №4, vol.7, pp. 4-17.

[2] Wang Z., (2011), The Application of Cloud Storage in the Library, Advanced Materials Research, DOI 10.4028/www.scientific.net/AMR.267.314.

[3] Mustafee, N. & Bessis, N., (2015), The Internet of Things: shaping the new Internet space, Concurrency and Computation: Practice and Experience. DOI: 10.1002/cpe.3483.

[4] Mukhopadhyay, D., Sharma, M., Joshi, G. and Pagare, T., (2013), Experience of Developing a Meta-Semantic Search Engine, 2013 International Conference on Cloud & Ubiquitous Computing & Emerging Technologies, Pune, India, Maharashtra Institute of Technology, DOI: 10.1109/CUBE.2013.38.

[5] Ono, M., (2014), Service Science in top IT Vendors, The 11th International Conference on Service Systems and Service Management, Beijin, China, Beijing Friendship Hotel, DOI: 10.1109/ICSSSM.2014.6874115.

[6] Chinchor, N., Christel, M., and Ribarsky, W., (2010), Multimedia Analytics Introduction, IEEE Computer Graphics and Applications, DOI:10.1109/MCG.2010.87.

[7] Chinchor, N., Thomas, J., Wong, P and Christel, M, (2010), Multimedia Analysis plus Visual Analytics = Multimedia Analytics, Computer Graphics and Applications, IEEE, DOI:10.1109/MCG.2010.92.

[8] Rongrong Ji, Yue Gao, Wei Liu, Xing Xie, Qi Tian, and Xuelong Li, (2015). When location meets social multimedia: A survey on vision-based recognition and mining for geo-social multimedia analytics. ACM Trans. Intell. Syst. Technol. 6, 1, Article 1 (March 2015), 18 pages. DOI: http://dx.doi.org/10.1145/2597181.

[9] Lieberman H. & Selker T., (2003), Agents for the user interface, Handbook of Agent Technology (pp. 1-21).

[10] Onykij B., 2010. Scientific and Technical Report on the implementation of Stage 5 of the State contract № 16.740.11.0129 dated 02 September 2010.

[11] Artamonov A. & Tretyakov E. (2015), «Automated proceeding of Big data for thematic thesauruses creation, International scientific and technological cooperation (p.200), Moscow, Russia: National Research Nuclear University “MEPhI”.

[12] Pasche, E.; Gobeill, J.; Teodoro, D.; Gaudinat, A.; Vishnyakova, D.; Lovis, C. and Ruch, P.; (2012), An advanced search engine for patent analytics in medicinal chemistry, Studies in health technology and informatics. 2012, vol. 180, p. 204-9, DOI: 10.3233/978-1-61499-101-4-204.

[13] Savenkov, D., Braslavski, P. and Lebedev, M., (2011), Search Snippet Evaluation at Yandex: Lessons Learned and Future Directions, 2nd International Conference of the Cross-Language Evaluation Forum, CLEF 2011, Moscow, Russia: Yandex.

[14] Dzemyda, G., Marcinkevicius, V. and Medvedev, V., (2011), Large-Scale Multidimensional Data Visualization: A Web Service for Data Mining, Springer Berlin Heidelberg, DOI: 10.1007/978-3-642-24755-2_2.

[15] Efthimiadis, E.; (2000), Data visualization in information retrieval and data mining, Proceedings of the ASIS Annual Meeting, v37 pp 444-446, ISSN: ISSN-0160-0044.

[16] Buys J. (2015), A New Search Engine Built from Open Source, Retrieved May 16, 2015, from http://ostatic.com/blog/duckduckgo-a-new-search-engine-built-from-open-source.

[17] Vaughan, L. & Chen, Y., (2015), «Data Mining From Web Search Queries: A Comparison of Google Trends and Baidu Index», Journal of the Association for Information Science and Technology Volume 66, Issue 1, pp. 13–22, DOI: 10.1002/asi.23201.

[18] Artamonov A. & Kshnyakov D. (2015), Information codification systems application for primary data resource cluster development, International scientific and technological cooperation (p.199), Moscow, Russia: National Research Nuclear University “MEPhI”.

[19] Paparrizos, I., Jeung, H. and Aberer, K., (2011), Advanced Search, Visualization and Tagging of Sensor Metadata, 2013 IEEE 29th International Conference on Data Engineering (ICDE) (pp: 1356-1359), Hannover, Germany;

[20] Hamlen K., Kantarcioglu M., Khan L. and Thuraisingham B., (2010), Security issues for cloud computing, International Journal of Information Security and Privacy, http://dx.doi.org/10.4018/jisp.2010040103.