A Three-tier Model for Structuring of Scientific and Technical Information

Alexey Anatolievich Artamonov, Anastasia Gennadievna Ananieva
Evheniy Sergeyevich Tretyakov
Dmitry Olegovich Kshnyakov, Boris Nikolaevich Onykiy
Larisa Vladimirovna Pronicheva
National Research Nuclear University "MEPhI"
Kashirskoe Shosse, 31, Moscow, 115409, Russia

ABSTRACT: The use of the World Wide Web for information analytical support of scientific research and development encounters such modern world's issues in the scope of IT as Big Data Mining. This paper describes an agent-based technique of data search in the web that were developed at the National Research Nuclear University (NRNU) "Moscow Engineering and Physics Institute (MEPhI)", department of “Analysis of competitive systems” and implemented in the Multi-agent information and analytical system (MIAS). The experiments of the agent-based data search techniques and the MIAS have been carried out and the results are shared within this paper. By developing the MIAS, we support international standards for data sharing. Reasoning from this fact, we enable all services and people to bear down the most crucial issues of the 21st century as Big Data.

Subject Categories and Descriptors
H.3.1[ Content Analysis and Indexing] ; I.2.11[ Distributed Artificial Intelligence]: Multiagent systems; H.2.8 [Database Applications]: Data mining

General Terms: Chinese text mining, sentiment analysis

Keywords: Agent, Multi-Agent System, Pertinent, Data Mining, Information Retrieval

Received: 8 January 2016, Revised 7 February 2016, Accepted 19 February 2016

1. Introduction

The World Wide Web and other computer networks have fundamentally changed the information communication systems in the world. Today the accumulation processes and the use of information in the web are characterized by huge numbers. According to the International Data Corporation (IDC), the amount of data generated in 2012 was 2.8 zettabytes, and by 2020 it is expected to reach 40 zettabytes [1].

There is no doubt that the network information technology helps to solve a number of fundamental problems in science, technology and industrial production. However, the use of the web for information and analytical support of scientific research and development has encountered the problem of information dispersion, i.e. the emergence of a huge number of scientific and technical information published on the websites of research centers, universities, large laboratories and scientific technology companies, professional associations and, finally, just individual specialists in the global network.

These new sources of scientific and technical information are very attractive as such information appears with minimal delay [2], and any user is able to contact the authors of various articles.

However, these new sources have the following drawbacks:

1. Information is published on every website in compliance with its structure, which makes the retrieval process difficult;

2. The information owners put forward various conditions for obtaining the information.

3. The number of information sources is so large for each research thematic area that a single user can hardly look through the whole at a time.

4. There are other issues that are common for all types of information, namely, the determination of the original
source and information reliability for example information can be changed on purpose, and in some cases metadata, such as publication date, author, and source, can be intentionally distorted.

A number of questions related to such information may appear as follows:

• Is this an original source?
• Who is the author?
• Has it ever been edited?

However the problems mentioned above cannot be solved by searching engines, like Google, Yandex and Yahoo, which simply provide references, encyclopedia materials and information related with specific users' concerns. Besides these searching engines prioritize the average user. Specifically, the first page of the search results for natural-scientific and technological information often presents the websites for major visitors, whereas the sites for specialized minorities to access the latest news are, at most time, shown on the second page and further that have rarely been touched upon by the average user.

In order to find the relevant information, users have to use the special syntax of a search engine, through which it is possible to build complex search requests and get the most relevant results on the basis of properly selected keywords, and logic operations between them. The challenge of the search engine syntax development is a time-consuming process that requires constant practice, so an average web user does not use modern information retrieval systems to the full [3]. The keywords in an ordinary user search query normally consist of two to three words without logical operators. The information retrieval languages are not capable of transmitting the content of documents and the meaning of requests, so the information loss and noise are inevitable consequences of such information search.

The above user problems - to get access to professionally relevant information - are now formulated as a problem of Big Data by various authors; in the present work, this problem is understood as the inability of a professional user or even a small group of users to solve all tasks of receiving and processing scientific and technical data from the web.

As a result, the question – "Is it possible to find such information technology that would solve the problem of Big Data without burdening the professional user with a large amount of time and getting overall awareness of the subject activity area?" – arises.

Developing scientific databases, e.g. Web of Science, Scopus (USA), and VINITI Database RAS (Russia), which are major information centers/hubs, is one of the classic methods of compression, concentration and classification of scientific and technological information. These databases contain useful information about publications in their selected information sources, different statistical characteristics of publication processes in various fields; however they only contain the references of the full text to be published.

Furthermore it makes a user perform search operations. Large dimensionality of summarization tasks, the credibility of different scientific and technical information sources require these hubs for information source credibility evaluation, which may not coincide with the evaluation done by a particular user. In particular, the hub may analyze the information published in English only from English sources, and sources in other languages would not be considered without translation. Consequently, it can bring about a loss of news completeness on various thematic areas.

2. Methodology

Modern information retrieval systems, commonly based on full-text search, enable an access of a high-level classical relevance — matching the user request to the result of the search [4], [5]. Common characteristics of automated search engines are: completeness, accuracy, relevance and pertinence.

Nonetheless, the quality of information search is characterized not only by relevance, but by pertinence — matching search results to users’ needs (expectations) [6], [7].

The relevance may be quite different from pertinence, and these concepts are constantly confusing. Nowadays, ensuring pertinence is one of the main issues in various developing information systems. Users need a minimum set of keywords to obtain the necessary information, and the frequently correlation of these requests with the target information is available only in users’ minds.

The results of most search engines do not often meet the requirements of high pertinence. This is associated with the features of natural language, such as synonymy and homonymy [8], [9]. In addition, the user of an information retrieval system cannot always clearly and unambiguously articulate the set of keywords that may lead to the desired result.

Low pertinence is often attributed to the complexity of the information requests generation for full-text search. These difficulties are caused by the following reasons:

• The ignorance of a set of keywords that define the searched documents semantics;
• The lack of accepted and the well-established terminology for the subject area;
• The lack of sufficient experience and qualifications for the generation of search queries.
As for the scientific and technological areas, the first two reasons listed above can be neutralized due to the fact that such areas are characterized by the well-established terminology, and as a consequence, specialists working in these areas are familiar with it. Characteristics of completeness and accuracy have always been controversial in information dialog type retrieval systems, which have caused the inability to achieve both - high completeness and accuracy [10], [11], [12].

One thing to be noted is, in the description of scientific and technological activities, the most stable category is its thematic focus. This feature of scientific and technical information is used in the classic decimal classification, i.e. library systems.

The stability of the classification is attributed to the duration of life cycles of objects of scientific interests, areas within the expertise of researchers and engineers. Thematic sustainability of research and technical activities enables practitioners to develop professional subset of languages, with the help of which the objects of scientific research and design-technological activities will be described accurately and completely. In fact, it is a professional subset of natural languages.

Additionally, the terms of these languages have a high stability that can be used as keywords for automated information retrieval systems. In the formulation of search queries of automated systems, it is possible to follow the further steps: to display thematic concerns of the user through keywords from thematic thesaurus, to specify a network source containing desired thematic information, and to continue scanning the source and process of thematic information without user intervention, and send retrieved and structured information by e-mail.

So information retrieval tasks in the web will be performed not by a user, but its representative - agent. The obvious advantage of this approach is that the agent is able to work round the clock and to automatically carry out the amount of work that is greater than the classic dialog mode search and to perform information structuring.

The definition of agent was formulated at the Tokyo meeting of the Foundation for Intelligent Physical Agents (FIPA) held in 1996: “an Agent is an entity that is situated in some environment, from which it receives data representing events occurring in the environment, interprets them and executes the commands that affect the environment” [13].

Nevertheless, in our case, the agent is a search software or a search robot that can be independently activated in the system on a specified schedule or by a user request and it performs actions either in the web or within agent-based systems in compliance with the predefined regulations.

Unlike the impersonal user concepts in the global information system, users of scientific and technical information can be classified as follows:

- A collective of users:
  1. Large polythematic scientific organizations and universities,
  2. Units (e.g. laboratory, department, and research group).
- Personal users
  3. Individual experts or small research groups consisting of 2-6 people.

Each of these three user groups has a role in science, information focus, and quality characteristics of information services.

The first group of users is interested in the completeness of incoming scientific and technical information from all global sources. Agent technologies solve this problem by overall automatically scanning [14], [15], [16].

The second group of users is interested not only in completeness but also in the accuracy of information, which is provided by agent technologies at high level by automatically filtering all the incoming information flow.

The quality of personal user service is characterized by the pertinence of the received information, and this rarely used feature reflects the degree of particular user satisfaction through the data received from the system.

Thus, it can be said that correspondence exists between the type of users and the type of agents to ensure the quality of their information service. The following are the three types of agents performing different functions in agent-based search engines:

- Agent-postman - an agent that interacts with web information sources on schedule and aggregates the information in a text database.
- Agent-courier - an agent that interacts with the agent-postman and the text database and organizes information under relevant thematic areas.
- Agent-referent – an agent that interacts with the user and the agent-courier in order to deliver pertinent information from the system database to the user.

The work overview of agent-based systems and technologies shows that they have achieved widespread practical use in areas such as unmanned moving objects and robots with different functions. The second area includes problems of various business information regarding decision-making systems. The following direction of the agent practical application is based on solving complex information retrieval tasks within industrial corporations that implement and support the lifecycle of large technical systems, such as the aircraft, shipbuilding and reactor
industries.

Since 2010, the authors have been working on the development of dedicated specialist support system in natural-scientific and technological directions. The developed system is called "Multiagent Information Analysis System (MIAS)" on scientific and technological areas, and it is put into pilot operation in a number of scientific Russian organizations. To solve the problem of maximizing the characteristics of completeness, accuracy and pertinence, the authors have developed a three-tier model of collecting and structuring of scientific and technical information [17], three-tier model of agent based search and structuring of STI, which includes:

• Gathering information from external network sources;
• Filtering and thematic categorization;
• Secondary filtering by personal dictionaries.

2.1 The first tier – Collection of information
On the first tier, the system solves the problem of regular collection of information from online sources available. Agents-postmen are the solution to this problem.

This type of agents is aimed at interacting with the environment and delivering the information to MIAS respectively to agents-couriers guided by the requirements of the search routing database.

The scheme of the agent system’s first level is presented in Figure 1.

Agent search database management contains information, such as the source name, primary language, status, date of entry, date of first launch, date of last response, the number of information messages, and link to the page with required news articles. Each agent is built individually for different sites and this provides a "pure" collection of data. Specifically, "pure information gathering" refers to the precise removal of information from the html page, cleaning text, double spaces, paragraph marks, tabs and so on. The configuration of one agent to a site takes half-an-hour specialist work or more (the agent building and testing are essentially included).

As the result of the configuration, the agent is passed to the stage of test operation, after which improvements are provided, and is placed into commercial operation. Currently more than 100 agents are working in the system on the thematic areas.

Setting a thematic route database includes an information sources search that is relevant to a certain topic. The operator searches for information sources, using keywords that describe the topic, search engine, and public and private knowledge bases [18], [19].

The task of the operator is not only to find the source of information but also to evaluate its credibility.

One approach for assessing the information source may be the source monitoring process and the evaluation of

Figure 1. Scheme of the first level
information published during a certain period; however this approach takes a significant amount of time. In the MIAS system, the following operational approaches for determining the credibility of information source are adopted:

- Domain assessment;
- Advertising evaluation;
- Structure evaluation;
- Evaluation of information supply services;
- Content assessment.

Domain assessment is performed in compliance with the parameters as follows:

1. Identifying the domain owner (can be done with the help of the service provided by the web site – https://www.reg.ru/whois/);
2. Determining the cost of the domain (can be achieved by means of such services as domain.ucoz.com);
3. Determining the age of the domain (can be carried out through using such databases as “Archive Wayback Machine”);
4. Identifying the domain level.

The credibility of an information source can also be identified through the use of content analysis. Practice shows that authoritative sources do not demonstrate inappropriate advertising, bright shimmering banners or pop-up windows; instead, they often have a very user-friendly interface with a logically correct section. The operator also analyzes the content of the information and determines the style of the text, the reference to the original source and the relevance of the information.

At the site, rapid assessment operator evaluates the number of services. These services can be provided in the form of RSS feeds, social networks and e-mail newsletters. Since the support of various information services requires major sources, an information source with a large number of information services can be considered as authoritative.

2.2 The Second tier - Filtering and Thematic Categorization
The second tier solves the problem of thematic categorization of information messages brought by the agent-postmen.

The stream of news reports from agent-postmen is uploaded to the full text database. Figure 2 shows the second level of agent-based model of search and structuring STI. The full-text database is processed by agent-couriers that distribute information among thematic directions. On this stage a service (abstract) information is formed and generated automatically by the same robot courier by highlighting the sentences containing the keywords of...
Agent-couriers are guided by the database of thematic thesaurus containing thesauruses that describe different thematic directions. To assign category to news message, it should contain at least one term from the corresponding thesaurus.

Within articles, terms search is performed by using a formal language of text search – “regex”, and thereby high accuracy is provided.

Thematic thesaurus can be written in several languages and it includes definitions for each term. The preparation of the thematic thesaurus requires the involvement of experts from the relevant scientific fields, who are working in dialog with translators. Such thesauri are presented in international format TMX 1.4 b [20], which is widely used in the software for the automated translation.

In the absence of relevant experts on the thematic areas, the operator has the ability to adhere to reference databases such as Web Of Science and Scopus. The operator forms the query to the database relevant to this scientific field and selects the conference data from the last year. The choice of conferences is not made on a random basis due to the fact that the conference data contains the latest recent research and data compilations.

In abstract databases “keywords” are attached to each entity of “conference proceedings” type to configure a new scientific direction in the system. Based on the counting of the most common terms, it is possible to determine the most popular technology trends. However, this algorithm simply provides a list of relevant terms, and further development of the thesaurus includes the selection of definitions of the respective thematic focus as well as the translation into several languages.

### 2.3 The third tier - secondary filtering by personal dictionary

Personalized information processing is presented at the third level (see Figure 3). A list of objects meeting user professional objectives and the personal thesaurus of each expert are used as reference materials.

The agent-referent is controlled by personal user dictionaries, on the basis of which automatic generation of annotations and the selection of facts to each message are performed. In this context, a fact is the sentence that contains at least one word from each dictionary.

Personal dictionaries of keywords allow any crossing of thematic interests. Configuring and compiling such dictionaries are carried out in a dialogue mode with the user. There are several types of personal dictionaries:
• Dictionaries that describe the thematic area, in which a particular user is interested;

• Dictionaries that describe the object;

• Dictionaries that describe the object phase, e.g. life cycle.

For example, it has been revealed that some user is interested in the research and development of the “Quantum technologies” area. In this case, the keywords describing the topic will be: “laser”, “quantum optics”, “free electron laser”, and “nuclear pumping”, etc.; keywords describing the entity: “project”, “research”, and “development”, etc.; keywords describing the stage: “developed”, “researched”, and “finished”, etc.

3. Results and Discussion

The authors carried out experiments to assess completeness, accuracy and pertinence of information after the second and third system tiers. All information stored in the system can be divided into four parts while the search request is carried out:

\[ N = A + B + C + D \]  \hspace{1cm} (1)

Specifically, “N” refers to the initial set of documents to be processed; “A” means information pertinent to the request; “B” is information pertinent not issued to a user request; “C” correlates information that is not pertinent to the user query; “D” refers to non-issued and non-pertinent information to a user request.

Completeness of information retrieval \( R \) is determined by the ratio of the number of found pertinent documents and the total number of pertinent documents \( (A+B) \), presented in the original set of documents:

\[ R = \frac{A}{(A + B)} \]  \hspace{1cm} (2)

The accuracy of information retrieval \( P \) is determined by the ratio of the number of found pertinent documents and the total number of documents \( (A+C) \) issued on the user request:

\[ P = \frac{A}{(A + C)} \]  \hspace{1cm} (3)

The presence of taken at the user request pertinent documents is called noise in information system. The ratio of information noise \( K \), respectively, is determined by the ratio of the not pertinent documents number for the user response to the total number of documents \( (A+C) \) for the

<table>
<thead>
<tr>
<th>MIAS Second-tier</th>
<th>Quantum technology</th>
<th>Biohybrid technology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indicator name</td>
<td>One week</td>
<td>One week</td>
</tr>
<tr>
<td>1. All information stored in the system (N)</td>
<td>163</td>
<td>303</td>
</tr>
<tr>
<td>2. Given out information pertinent to the request (A)</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>3. Information not given out and pertinent to the request (B)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4. Given out information not pertinent to the request (C)</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>5. Information not given out and not pertinent to the request (D)</td>
<td>157</td>
<td>293</td>
</tr>
<tr>
<td>6. Completeness of information retrieval (R)</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>7. Accuracy of information retrieval (P)</td>
<td>0.5</td>
<td>0.4</td>
</tr>
<tr>
<td>8. Ratio of data noise (K)</td>
<td>0.5</td>
<td>0.6</td>
</tr>
</tbody>
</table>

Table 1. Results of experiments on the MIAS second-tier testing
user request:

\[ K = \frac{C}{(A + C)} \]  \hspace{1cm} (4)

The experiments were conducted in two thematic directions – “Quantum technologies” and “Biohybrid technologies”. Experts on the thematic direction have evaluated the messages pertinence. Table 1 presents the results of the experiment on the assessment of completeness and accuracy of the MIAS second-tier performance through periods on a weekly and monthly basis.

Table 1. Results of experiments on the MIAS third-tier testing

<table>
<thead>
<tr>
<th>Indicator name</th>
<th>Quantum technology</th>
<th>Biohybrid technology</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>One week</td>
<td>One month</td>
</tr>
<tr>
<td>1. All information stored in the system (N)</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>2. Given out information pertinent to the request (A)</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>3. Information not given out and pertinent to the request (Å)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4. Given out information not pertinent to the request (Ñ)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5. Information not given out and not pertinent to the request (D)</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>6. Completeness of information retrieval (R)</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>7. Accuracy of information retrieval (P)</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>8. Ratio of data noise (K)</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 2. Results of experiments on the MIAS third-tier testing

As described in Table 1, the completeness of information search is achieved on the second tier of the system, while the precision of information retrieval for a particular specialist is at the level of 50%. It should be noted that the number of informational messages is reduced by 30 times respectively.

Table 2 shows the results of the MIAS third-tier experiment on the assessment of data completeness and accuracy, considering the fact that the third tier starts with the results of the second.

As described in the above table, the use of personal user dictionaries has managed to achieve almost 100 % of completeness and accuracy of information retrieval in the MIAS. Under the thematic “Biohybrid technologies” one message has not been given out but is pertinent to the user. This result stems from the fact that the user dictionary includes the word “self-assembling” whereas in the identified article was the word – “self-assemble”. With the results of the experiment there is adjusted custom dictionary, and the second experiment showed 100% of completeness and accuracy.

In turn, the experiments on time evaluation of categorization of the information stored in the database were carried out, and it is important as after each launch of the agent-postmen there comes new rubrication process for all stored in the system thesauri by means SQL queries. Positioning 58 search agents, which for all time of system operation collected 5607 news reports, on April 1, 2016, the process of categorization took 6 minutes 41.3 seconds.

Streams are used for speeding up the message-process ing algorithm; a series of experiments, from which the results have identified the optimal number of streams for a certain number of information messages, were conducted.

250 informational messages, which in turn were processed with different number of streams, were randomly selected. The results of the experiment are presented in Table 3 and Figure 4.

The results of this experiment revealed that the optimal algorithm is when 1 stream is used for 25 messages.

The results of experiments on the using agent-based technologies for information and analytical support of specialists working in natural-scientific and technological environment are shown as follows:

- The speed of collecting and processing scientific and technical information from the Internet quadruples in comparison with the normal dialog mode.
A three-tiered approach allows us to improve the basic characteristics of information retrieval systems quality – completeness and accuracy.

3. Conclusion

According to the results of the experiments, it can be concluded that the use of multi-agent systems in the information support of research organizations, research teams and individual specialists allows to collect pertinent data from multiple data sources and to structure them in compliance with the database model adopted for performing a given task with high degree of efficiency.

The indicator of information sources coverage completeness is determined by the completeness of the routing database that updates dynamically.

High accuracy in servicing both collective and personal users is provided by creating and maintaining two management databases: database of thematic thesauri and database of the user personal dictionaries.

One of the most labor-intensive works requiring highly-skilled performers is a compilation of thematic thesauri. This work should be implemented in parallel with different organizations around the world adhering to international standard TMX 1.4 b and the results of performed work in different national languages should be shared.

Acknowledgements

This work was supported by the Competitiveness Program of NRNU "MEPhI".

References


