Semantic Structuring of the Digital Library Content

I.A. Filozova
e-mail: fia@jinr.ru, Laboratory of Information Technologies, JINR, Dubna

Introduction

The volume of information processed is increasing at a rapid pace today [1,2,3]. It is promoted by global information system Web, modern processing technologies of dynamic information, software tools for creation effective information technologies search and process information, and the spread of corporative and global networks. Under these conditions, the demand for retrieval systems, data analysis, problem solving under conditions of uncertainties is sharply increased. It's important to have an effective tool for the study of information in arrays of scientific publications as the main product of scientists and researchers for Specialist in given field of knowledge. Publications in a scientific style have an important feature: they form a different by character semantic relations with other publications. There are relations formed by quoting, links of scientific inference, the usage relations; relations of opinions and professional judgment, relations between staff and organizations, and the relations, reflecting the logic of presentation the author's thoughts within this publication, the themes, the subject area. Structuring and support these relationships can provide new opportunities to study a set of documents digital libraries (DL), thus improving their quality.

1. Problematic Situation

Main goal of the information search for the user is a satisfaction of his information need. It's possible by two ways: by the information request (establish an information request within the specified information search language; by question (ask the question in the subject domain language - truncated natural language).

The main criteria for estimation of the effectiveness of the search engines — the speed, accuracy and completeness of the answers. Accuracy is determined by which of the information issued in response to a request is relevant. Completeness is characterized by relation between all relevant information available in the database, and that part included in the response. Besides the estimation of search engines takes into account what types is supported by system, what is form for presentation of the search results and what level of user training is required to work in this system.

Search results depend of that how well the user has formulated him request/question. Often the user doesn't understand well if the received response is the pertinent to the him question. Pertinence indicator is a pragmatic metrics, it describes correspondence found documents to the information user need, regardless of how fully and exactly this information need is expressed in the information request [5]. The pragmatic level of the information search requires the additional knowledge about subject domain. Factors of achievement the insufficient level pertinence are: inadequate user knowledge about subject domain; unclear vision of what should be the search result; ineffective search mechanisms.

So, the creation of effective mechanisms to search the answers to the questions in the digital information collections is the actual problem. But it's necessary the semantic structuring of the content.

Also research results, scientific and engineering efforts represented in publications have semantic relations between them via citation mechanism. The linkages between members staff — the author of publications, their affiliation, participation in the collaborations, experiments, projects, etc. Description of these relationships and their properties opens up new possibilities for studying of a documents corpus of digital libraries. I'll say about relations, reflecting the presentation logic of the author's thoughts in this publication, topic, subject area. Discussed approach is based on the technology of the structuring of scientific texts by the logic-semantic network (LSN) Question-Answer-Reaction (QAR) for the organization of semantic search in digital libraries [4,5].

Development of methods for the organization of information retrieval confirms scientific and practical interest in solving this problem.

2. Approaches to Solving the Problematic Situation

Traditional approaches to information retrieval can be divided into three groups: the methods of the index (or binary) search, statistical methods and methods based on knowledge bases. To fully satisfy the information user needs theories and methods of semantic web, content analysis and text mining are widely used in the search engines now. Development of methods of the organization of information retrieval confirms scientific and practical interest in solving this problem.

3. Logic-Semantic Network *QAR*

Any scientific and practical knowledge area includes the subject of study that may be presented problematic field (the list problematic questions), that is the basis for the scientific and practical activities. Problem questions may be presented in a hierarchical tree on the principle "from the general to the particular". Some questions have the possible alternative answers and ways to implement them (reactions) already. It's required a certain reaction to understand the question. Answers may give rise to questions in turn. Thus, the problem question is related to a particular subject domain and it is revealed by the semantic structure a question-answer-reaction that is open (variable) over time. The knowledge accumulated in a subject domain may be presented an open set of logical-semantic networks, ordered by subject topics. The problem of the subject domain may be formulated in the form of a question. Revealing in question of such meanings as the question theme, the question content, the question amount allows to find relevant LSN, that may contain the answers and needed explanations (reactions). Thus, the search and processing system based on LSN — another way to improve technology of the information processing.

3.1. Cognitive Function of the Question

Communication of the specialists in some subject area is effective when it occurs in the form of a question-answer. The question as a tool of information search is very high. It is a bridge between the known and unknown.

Cognitive function of the question is aimed to the supplement, refinement and development of the previously obtained general representations of objects and phenomena of reality. The process of asking question and the answer search is a complex iterative process [5]. The question is based on an already-known knowledge that acts as a datum question always. The answer search assumes to address to a specific area of theoretical or empirical knowledge that is called the answer search scope. Setting process of the adequacy question and answer is aimed at the detection of the possible inconsistencies in the answer. Based on that, answer search scope or datum question or subject research is expanded.

3.2. Basic Statements of the LSN Question-Answer-Reaction

Logic-semantic network Question-Answer-Reaction - a set of the questions, answers and relationships between them forming an uniform system. It means:

1. Set of questions and answers belongs to a particular subject domain;
2. Set of questions and answers on the principle of hierarchical ordering "from the general to the particular";
3. The questions are placed on an odd level of the hierarchy, the answers — on even level;
4. Questions i-th level of the hierarchy are related only and only with the answers i +1-level;
5. Questions i-level are related with the answers i-1 level;
6. Question i-level semantically are linked with the answers i+1-level if it satisfies the conditions A) or B). Upon satisfaction of the condition A) it's final vertex. Upon satisfaction of the condition B) questions i+2-level follows from this answer;
7. The questions that reveal many answers on i = 2 level partially or completely covering the topic subject area are placed at i = 1 level;
8. The questions that complete and clarify the answers i = 2 level are placed at i = 3rd level.

Question - query expressed in the interrogative sentence aimed at the development, refinement or supplement of the knowledge.

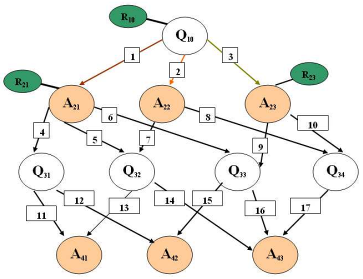
Answer - a realization the cognitive function of the question in the form of the new obtained judgement. Answer must be built in accordance with the content and structure of the asked question. Only in this case, the answer is regarded as relevant.

Reaction - a semantic description of the question and answer.

Types of reactions:

1. Question Reaction - a description of the datum question *(to understand the environment and causes of the question and to establish the semantic adequacy with the answer scope).*
2. Answer Reaction - a description of the answer scope *(to understand the question semantics and relationship with answer).*

Thus, QAR model may be presented by directed graph, where nodes are questions and answers (fig.1). Questions are placed on the odd level, answer — on the even level. Edges — the relations between them. Navigation is a motion way along *LSN*, controlled by the user.



№ edge	Way
1	1,4,11
2	1,4,12
3	1,5,13
5	1,5,14
6	1,6,15
7	1,6,16
8	2,7,13
9	2,7,14
10	2,8,17
11	3,9,15
12	3,9,16
13	3,10,17

Figure 1: LSN Graph Question-Answer-Reaction

The role of the reaction in this unit *QAR* is very important. It helps the user understand where the question (Datum Question) and answer (Search

Scope) appeared. Reaction can be text information, with links to primary sources, illustrative material (drawings, graphs, tables, slide shows, videos, etc.) and / or a combination thereof. Reactions help the user understand the semantic field questions and received answers to it, thus can improve pragmatic metrics of information search — pertinence Illustration of the reaction role is presented in [5].

System based on LSN is open (filled, variable in time). During the interaction with the system user can refine and expand LSN itself. So, user becomes the co-author of the semantic LSN space. This is the adaptation of the system.

3.3. Formal View of Subject Domain

Accumulated knowledge in the subject domain is expressed in the scientific reports, monographs, articles, educational materials, information collections, books, dictionaries, etc. It's possible to present the whole volume of information as the set of the ordered thematic sections, each of that reflects a certain aspect of the subject domain. Each topic can be associated with LSN Question-Answer-Reaction. Integrated semantic structuring of fund digital libraries on the LSN basis leads to the creation of multilevel network structure that can be the basis for the navigation mechanism. User has possibility to navigate in the horizontal and vertical directions (fig.2). Motion from i-th to the (i + k)-th level of network deepens the knowledge. Horizontal motion network expands the knowledge. Motion up the network summarizes knowledge.

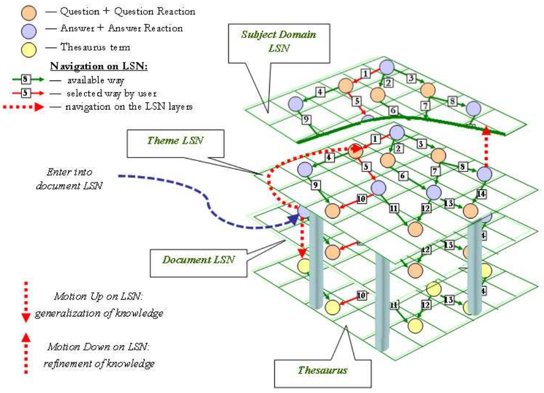

— Question + Question Reaction
— Answer + Answer Reaction
— Thesaurus term
Navigation on LSN:
8 — available way
5 — selected way by user
— navigation on the LSN layers
Subject Domain LSN
Theme LSN
Enter into document LSN
Document LSN
Motion Up on LSN: generalization of knowledge
Motion Down on LSN: refinement of knowledge
Thesaurus
1
2
3
4
5
6
7
8
9
10
11
12
13
14

Figure 2: Multilayer Related Set of Graphs

3.4. Methodical Basis of the Information Retrieval in the Digital Information Funds Based on LSN

3.4.1. Analysis of Scientific Texts

Before to start the creation LSN for the resource (storage unit of digital fund) there is need to make sure that it is a scientific text, i.e. text in the scientific style of speech. There are several varieties of the scientific text: academic; popular science; research and training. Regardless of the genre such texts have the basic characteristics: consistency of presentation, semantic accuracy, generalizability, objectivity. Scientific style has lexical, morphological, syntactic and structural features, as well as the typical use of the expression. For example, personal, subjective opinion, usage the pronoun I and verbs in the first person singular are not accepted in a scientific text. Indefinite-personal, impersonal, definitely-personal statements are preferred. In scientific language complex sentences are applied, because complex sentences with subordinate clauses express generalization, reveal a phenomenon, regularity. Usually scientific text includes the following components: title, information about the authors, abstract, keywords, the main text, bibliographical references, bibliographical list.

3.4.2. Analysis technique of the of scientific texts for the formation of LSN QAR

The technique is applicable only to scientific texts. To build document LSN it's needed to analyze it. The document is studied by the expert in terms of:

1) semantic matching the title and content;

2) set of filters:

Filter 1 (F1) – General Part. F1 includes an analysis of the problem, its history, overview, topicality.

Filter 2 (F2) – Author concept. F2 includes new terms introduced by the authors, traditional terms with the author's interpretation, the narrowing semantics.

Filter 3 (F3) – Examples and illustrations. To clarify difficult places in the text, reduce the text size under stringent restrictions on the volume.

Filter 4 (F4) – The idea of the author. Describes and explains the author's main idea.

3) The formation of the basic questions that are answered by the text.

On the obtained material LSN of the low level is based (LSN document):

1. Keywords are extracted from the name of scientific text.
2. Thesaurus is formed.
3. A hypotheses are proposed what problems are presented in the analyzed text.
4. The text is divided into several parts (information blocks) that are applied to filters F1, F2, F3, F4.
5. The basic idea is formulated for received information blocks — a set of proposals (statements).

Comparing the resulting statements with n a hypotheses we are able to make conclusions how the text title corresponds to its real content. After this step, it's possible to star the formation of LSN:

1. Formulation of questions to selected information units.
2. Selection the answers of the analyzed text, and links to them.
3. Formation reactions of questions and answers. For LNS scientific text (lower level) the reaction of questions and answers are generated from the information units on the filter F1 and bibliographic references.
4. Formation of the graph QAR.

As an optional service, the user may be offered some visualization for navigation.

4. Organization of Information Retrieval Based on LSN in Digital Funds

This approach is based on a qualitative analysis of scientific texts. So:

- structured, poorly structured information of subject domain maybe represented as a logical and semantic networks QAR;
- logical-semantic network reflects a certain subject domain;
- subject domain is represented by a set of themes;
- objective problem may be presented in the question form (or set of questions) ;
- solution may be represented in the answer form (or set of answers) ;
- solution way of the problem may be represented by a unified search mechanism of reactions on a set of logical and semantic networks;
- quality of the problem solution is represented as pertinence of answers to the question.

It is assumed that:

- LSN set are the basis for the structuring of arbitrary texts of scientific and technical information,
- LSN set are the basis of structuring the domain knowledge,
- search the relevant information requested may be made on a unified search engine based on LSN.

Thus, the creation of a semantic search engine based on the LSN *QAR* involves the following steps:

1. The development of theoretical propositions of technology for answers to questions for specialized scientific corps.
2. Development of automated technology for formation and support of specialized scientific corps.
3. Development of structural and functional model of the semantic search engine based on the LSN *QAR*.
4. Implementation of prototype system components.

Semantic search engines based on the LSN *QAR* may have a wide range of applicability, including in digital libraries. To implement such a system within a particular DL it's need:

1. Develop a LSN set;
2. Implement information search engine in mode answers to a question;
3. Implement a navigation mechanism for moving up (from the particular to the general) and down (from general to specific) at LSN.

Conclusion

From a user perspective, such a system allows to find an answer to the question in most cases. The user asks a question and gets an answer with additional information in the form of question and answer reactions that help to correct the question or to use refined or generalized question. In the question-answer mode calculation of the measures proximity between question asked of the user and question existed in the LSN already is implemented. If such a question is absent, it can be inputted into the system later.

Creation and support of such a system requires a great deal of serious work, both technological and organizational. Creation of a catalogue service is a time-consuming manual process. Therefore, technology of the creation and support of the catalogue on LSN basis requires maximum automation to provide workstations for analysts who will be engaged in the formation of LSN documents and subject domains. With the successful implementation of this system new opportunity will be provided to users of digital library — get answers to questions asked in natural language.

References

[1] World's Technological Capacity to Store, Communicate, and Compute Information//Science 1 April 2011: Vol.332, no. 6025.- pp. 60-65. - DOI: 10.1126/science.1200970

[2] Registry of Open Access Repositories. URL: http://roar.eprints.org/

[3] OpenDOAR - Directory of Open Access Repositories. URL: http://www.opendoar.org/

[4] V.N. Dobrynin, I.A. Filozova. The search based on the logical semantic network "Question-Answer-Reaction" – Proceedings of XII Russian Conference RCDL'2010, Kazan, Russia, Oct.13-17, 2010. – Kazan: Kazan State University, 2010. – p. 301-308. (on russian)

[5] I.A. Filozova. Technology of semantic structuring of the digital library conten — Proceedings of 5th International Conference "Distributed Computing and Grid-technologies in Science and Education", LIT JINR, Dubna, Russia, 16-21 July, 2012 – Dubna: JINR, 2012. – p.117-122.