ON THE PROBLEM OF MODELING OF THE HORIZONTAL RELATIONS BETWEEN DOCUMENTS

Yu. V. Leonova, V. B. Barakhnin, A. M. Fedotov
Institute of computation technologies SB RAS, Novosibirsk, Russia
e-mail: juli@ict.nsc.ru, bar@ict.nsc.ru, fedotov@sbras.ru

Рассматриваются информационная модель горизонтальных отношений связей между документами, представляющими соответствующие сущности на основе бинарных отношений с дополнительными атрибутами, и ее применение для построения научных информационных систем.

Introduction

While creating the information systems with access via Internet for scientific community it is very important to give the user a maxim comfortable navigation on the net, using the possibility of hypertext in full measure. Unfortunately, these possibilities are not always realized. In particular, considering the information systems of mathematic direction in the best known elaborations of our country — portals Math-Tree [1] and Math-Net.ru [2] — the inner hierarchical connections between the documents are absent. In some foreign systems such connections are available but as a rule, these system are specialized on the genre of resource. For example, Portal MacTutor History of Mathematics [3] contains a very detailed information with cross-references but bibliographic information is represented as a short list of works.

In projecting the information systems a problem arises of the possible disagreement of the information. At first, including the information of heterogeneous points in the documents can result in the appearance of multiple information about the same object. Such situation is possible for example when a person works in different organizations, takes part in different projects and has many publications. This may cause serious problems in case of necessity of the appearance of different versions of the information as a result of its modification.

Moreover, to represent the complex documents when one document is a part of another (completely or partly), it is necessary to work up an approach to establish relations between the documents. Such a situation appears when a true saying may be constructed about the points described by documents (having interest from the point of view of the information systems contents), alike: "Point A is (was) something in relation to point B", or "Point A has (had) point B in some quality". For example: "Euclid — is the author of "Beginning"" or "S.L. Sobolev was the director of the Institute of Mathematics of SB RAS". It is well seen that the types of such relations may be different and this circumstance must be taken into account in the process of working up the model of relations between the documents.

Therefore, it becomes actual to work up the technology of identification, specification and visualization of horizontal relations between points, information of which is contained a number of documents, and also between the documents being part of complex documents. One of the main elements of this technology is working up the information model of relations and subject connections between the documents of the system.

Note that the Library systems constructed on the basis of the protocol Z39.50 and its versions [4] the complete doubling of official information is fulfilled. A similar situation appears in the information systems constructed on the basis of LDAP-directories [5] in which there is a powerful system of cross-references, but the model used does not admit relations "many-to-many". Even if such relations appear, there is a necessity to double information, which can result in a discoordination of the information in the system.

Therefore, in the information systems similar to one worked up by us it is advisable to keep information in the singular example, in necessary cases establishing relations "many-to-many".

Naturally, the methods for solution of such problem were discussed earlier in some works [6, 7]. However the main approach to representing data in these works is considering multiple relations with their following decomposition in the process of normalization. In contrast to them, we construct an information model using only binary relations attaching them additional attributes not fitting the common scheme.

Therefore, the decomposition is carried out on a higher level of abstractness from the data structure, which makes our model more universal.

1. Model of the document in the system

Information system is represented by a multitude of documents connected by different relations, describing some points (i. e. objects, facts or conceptions). The information about some points is kept in the system either in the form of the documents representing, describing or modeling it or in the form of mentioning of this point, present in the other documents (contain indirect information about this point).

Obligatory attribute of information system distinguishing it from usual web-sites is the presence of a catalogue containing the metadata of the documents.

According to the standards of open systems interconnection (OSI) [8], the structure and contents of the documents must be described in line with international schemes of data. For describing the corresponding schemes of data they use metadata determining the structure and the semantic content of the document. In our system the information resource equipped with metadescription (metadata) in consequence with recommendations of OSI is called the document.

Determination 1. Document d_i is called a couple:

$d_i = <S_i, V_i>$,

where S_i — structure of the document in the consequence with selected data system;

V_i — content of the document (information filing).

Determination 2. Collection — multitude of documents with definite fixed structure, the content of which is similar to thematic direction.

From the point of view of unification of the work with the documents, we shall represent the information system in the form of a set of collections. Metadata describing the structure and the content of the documents are subdivided into descriptive and structural.

The structural metadata determine the structure and the properties of the documents

according to which they are processed (types, connections, formats of representation, limits to access direction etc). The descriptive metadata describe the semantic content of the document (title, short content, etc). Note that the descriptive metadata characterizing the documents may be part of the document and at the same time may contain information of the document (main and additional, such as authors, title, date of creation etc) according to selected data scheme.

The element of scheme data of the present collection will be called a structural element (further, an element), has an identifier and some properties.

Therefore Element E is a totality <ID, P>,

where ID — identifier of the element;

P — properties of the element.

An example of the element has a value (or content). The properties of the element determine the character of the work with the element.

The element has type selected from a dictionary. The type determines the rules of the work with the element and, therefore, is a property of the element.

Examples of the element: title of the document, abstract of the document, surname in the visiting-card, authors of the document. Value of the element is its concrete content part and properties of the element describe its structure. For the element of the visiting-card "surname" value is Matveev, identifier — 1, properties — "word" type.

Note that an external object may be the value of the element also. For example for the element of the visiting-card "photography" the value is the external object — graphic file.

Structure of the documents is a set of structural elements. The content of the document is joining up the values of the elements forming the document.

2. Information model of information system

In accordance with the principles of OSI [8], documents must be transferred inside the information system in a unified exchange format. By an exchange format we call such a format which contains both the description data structure and the data. Therefore, the information system must have at least 3 level presentation of the information which corresponds to the principles of modern technologies of construction of information systems (Microsoft.Net [9], XForms [10]). The basis of such conception there is a principle of separation of storage system from structure and presentation of the information (fig. 1).

• Level of storage of information — stored representation of data and metadata in a form determined by the repository of data.

• Level of processing of information — internal (exchange) representation of the document. Document in exchange representation contains data and structure.

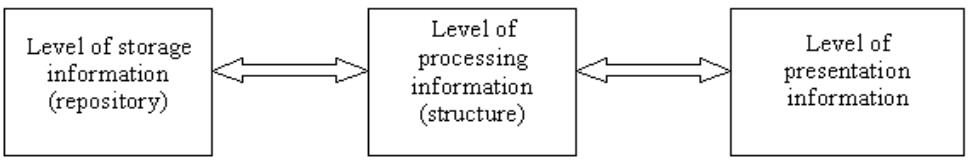

Fig. 1. Scheme of movement of the document in the information system.

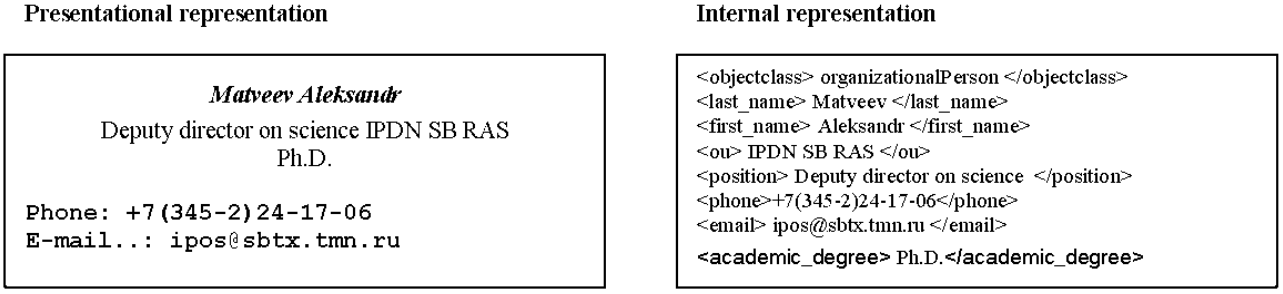

Fig. 2. Representation of the document.

• Level of presentation of information — representation of the document in the form comfortable for a user.

The movement of the document between the levels occurs to both sides, which gives a possibility of both presentation of the information and adding the changes.

In the multi-user information system it is necessary to supply: 1) differentiation of rights to the access of the user to information resources; 2) differentiation of rights to access to structural elements of the document. Solution of these problems may demand an introduction of two additional levels, but they are beyond this article, as we discuss only the technologic part of the system.

Consider the functioning of the system in the direction of movement of the information "level of storage — level of presentation".

1. "Level of storage" determines in what way the system stores the document — in relational table, directory (LDAP) etc. Level of storage performs the selection from the repository and transfers it to the level of representation for working up.

2. On the "level of processing" the calculation of relations and generation of internal representation of the document from the stored representation is performed. Information of the level of representation is necessary for supplying interoperability with other systems. Internal representation is an intermediate document containing a set of elements in an exchange format (associative array containing data and meta description), for example XML. Internal representation is transferred to the level of presentation or external inquiry.

3. On the "level of presentation" a presentational representation of the document is generated from internal representation by using a template.

Figure 2 shows an example of presentational and internal representation of the document.

3. Model of relations between the documents in the information system

To solve the problems formulated, we must determine the connections (relations) between the documents. As the basis of our model of relation between the documents in the information system we took the model RDF [11] describing presentational representation the resources and relations between them. Description of the resource in the RDF is a totality of statements about the properties of the resource. Every statement represents: resource, named property, and its meaning. The relations between the resources are represented by named properties.

For example, a statement that Korobeinikov S.N. is the author of the book "Unlinear deformation of solid bodies" in notation of N-Triples (RDF) may be expressed in such way:

<Korobeinikov S.N.>
<AuthorOf>
<Unlinear deformation of solid bodies>.

The main difference of our model from RDF is that the relations constructed by us are transferred on the level of the elements, determining the structure of the documents. In the considered information system the relation itself is determined the structure not by resource but by the structural metadata of collections documents of the system.

In our system the relations between the documents are established specifying on a number of documents of binary relations which, according to one form of notations used by RDF, may be written as A(R,V): object R has attribute A with value V. For example, the fact that Barakhnin V.B. holds a post in the Institute of Computation Technologies SB RAS is written down as follows:

Post ("ICT SB RAS", "Barakhnin V.B."),

where Post — some value from the list (thesaurus) of posts.

In the information system for scientific association we distinguish 2 types of relations.

- Relation of the order between the documents constructing the hierarchy of subordination in the collection, for example relation of hierarchy of subordination between the documents in the collection "Organizations":
  Head("Chair of mathematic modeling", "MMF of NSU").
  Note that such type of relations assumes the establishment of only one-way connection between the documents.

- Relations of connection between the documents, for example relation of the type of belonging between the documents of collection "Organizations" and documents of collection "Persons":
  Post ("ICT SB RAS", "Barakhnin V.B.").
  This type of relations admits the establishment of 2-way connection between the documents what means that the reverse connection may exist concurrently, for example Position ("Barakhnin V.B.", "ICT SB RAS").

Direction of connection is determined by the order of the arguments in relation A(R,V). Therefore any object may also play a role of the value.

The difference between the relations of the first and second types is that the relations of the first type have initially a property of hierarchy and the relation of the second type has no initial properties. The properties of the second type relations are determined for every concrete relation. As a rule, the relation of the first type has not more than one attribute, for example type of subordination (territorial, scientific-methodical etc).

Relations of the second type usually have some additional attributes. For example, relation of the type "Post" does not simply describe the belonging of a person to an organization but also has the next attributes: name of the post, keywords, date of assignment, date of release etc.

For the relation A(R,V) argument R will be called a head document and argument V — a subordinate document.

A document in the system may be connected with any number of documents. Between two documents there may be direct and reverse relations.

Direct relation — relation of the head document to subordinate, for example relation of the document "visit card of organization" to the document containing multitude of subdivisions,

employees or list of additional information. Document from collection "Persons" or "Organizations", may be connected with the documents from the collection of additional information, for example, list of additional data.

Reverse relation — relation of the subordinate document to the head document.

For the one-way relations parental document always knows its daughter documents and daughter document knows nothing about its parent. For supplying the navigation in collections it is necessary to have a registration of reverse relations of the document.

Distinguishing 2 types of relations between the documents, we solve 2 problems:

— navigation in collections;

— establishment of connections between the documents (hyperreferences, insertions).

Based on the relations of the second type, it is possible to distinguish 2 types of elements in the document:

— elements, the content of which does not depend on the values of relations attributes;

— elements, the content of which can depend on the values of relation attributes (for example, official information depends on the post of a person in organization).

Note that the elements of second type can contain lists of references on other documents, lists of insertions.

Figure 3 shows direct connections of collection "Organizations" and collection "Persons", fig. 4 — collection "Publications" and collection "Persons", fig. 5 — structure of additional attributes of relations of second type taking into account the multiple type.

However, using the considered scheme does not solve all the problem which appear while constructing information systems for scientific community. For example, the problem of the loss of actuality of the information concentrated around organizations, associations etc.

We may be interested in the method of the solutions of operation equations of Bubnov-Galerkin or biography of I.G. Bubnov, but it is unlikely that we will look for this information by searching the data of the Navy Academy where Bubnov worked.

Therefore information should be grouped in such way:

— around scientist persons;

— around conceptions and facts of science;

— around realities of surrounding world described by science (for natural and some humanitarian sciences).

Within the scope this article we are interested in the first approach (technique of constructing information models within the approach with the use of thesauruses considered in [12]; the third approach is based mainly on the systematization of the subject considered within the scope of corresponding science).

For the information model of a system the basis of which is a person, in searching a necessity appear to compare all its positions (also concerning publications) i. e. to use the approach reverse to that described above. Solving of this problem with the help of context inquiries

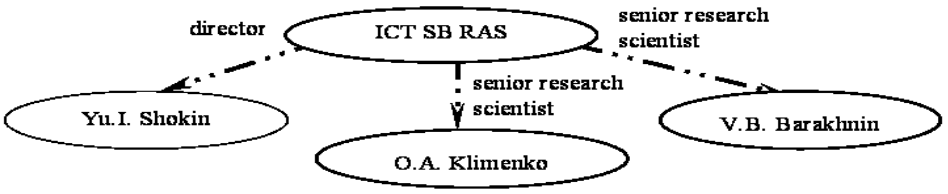

Fig. 3. Connections between documents of collection "Organizations" and collection "Persons".

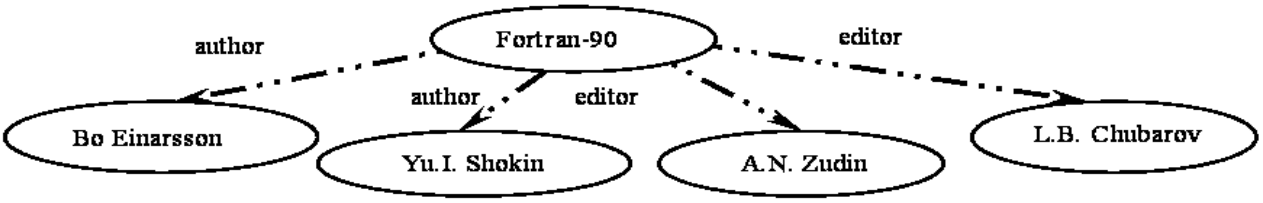

Fig. 4. Connections between documents of collection "Publications" and collection "Persons".

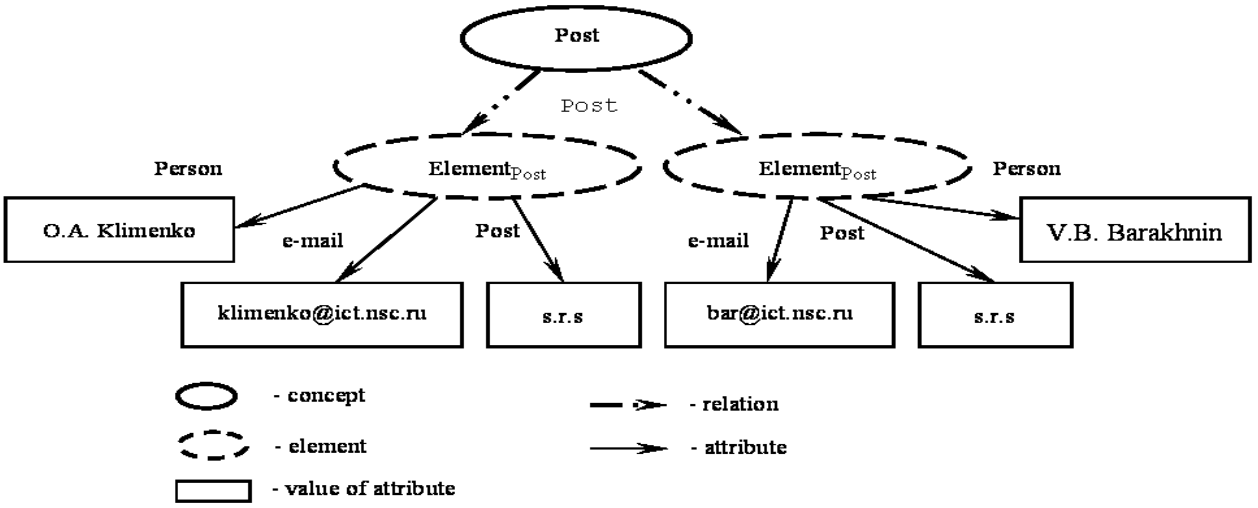

Fig. 5. RDF-representation of additional attributes of relations of second type taking into account the multiple type.

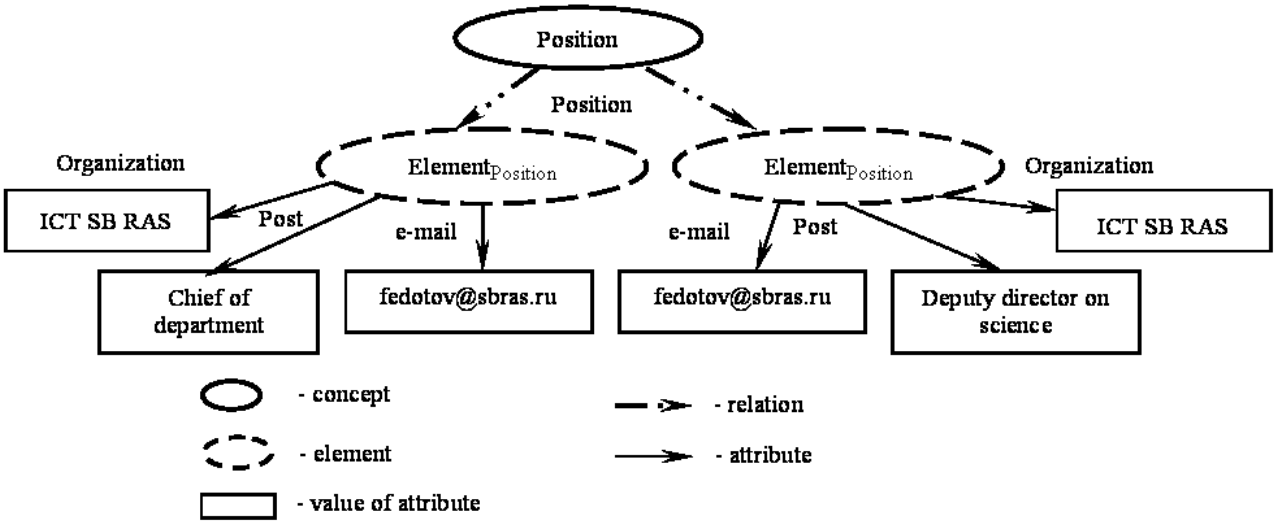

Fig. 6. RDF-representation of additional attributes of relations of second type taking into account the multiple type.

(even to concrete field) is not always comfortable as it may result in irrelevant documents. Therefore a necessity appears construct a reverse model of relations, which should have a universal character.

Therefore, the documents of information system are grouped according to the following principle: there is a specially distinguished collection "Persons" and multitude other collections: "Publications", "Organizations", "Associations" (societies, journals, etc) and all relations are constructed around persons.

A person may take different positions, be an author or the editor of the publication, take some post in organization, be the chairman or member of the chair, etc. All these cases are represented by one type of relations — "Position", which can get different names (director, post-graduate, chairman of the chair, author, etc).

Figure 6 shows RDF-scheme describing representation of multiple element "Position" from scheme of data of person collection, containing the Post of the person. Person A.M. Fedotov is connected with organization ICT SB RAS with relations "Deputy director on science" and "Chief of department".

Based on the documents connected with relations, we can from inquiry generate the following representation of the document:

```
<id = "14">
<LastName>Fedotov</LastName>
<FirstName>Anatolyi</FirstName>
<MiddleName>Mikhailovich</MiddleName>
<Position>
<organization name="ICT SB RAS">
<post>Chief of department</post>
<post>Deputy director on science</post>
</organization>
</Position>
```

4. Use of the model in the information system "Web-resources of the mathematical content"

The information system "Web-resources of the mathematical content" [13, 14] created in ICT SB RAS is intended to catalogue mathematic Internet-resources with the aim to supply the needed information to relevant searching. In the process of the work with the system it became clear that the tree-like structure of the information, regulating the documents according to their type (person, society, institute, department, laboratory, group, faculty, chair, scientific school, conference, seminar, edition, journal, book, article, project, package of programs, library, collection of data base, forum) and also in the correspondence with "Mathematic Subject Classification" [15] used by American and European mathematic societies, is not enough. In the new version of the system we envisage the establishment of internal connections between the documents in the correspondence with to described technique. These connections permit at output of information about organization to depict in some way information about persons working in it and publications of these persons and, when giving information about the person, to give the a hyper-reference on the site of corresponding organization and list of resources — publications of the person etc. On the first step mentioned modification touches the resources of the issue "Mathematicians of SB RAS" [16].

Information of the issue may be divided on two parts: biographic and bibliographic. Biographic part is based on the information system "Database of organization and employees of SB RAS" [17]: list of organizations and employees and also information about relations between the elements of the named lists and attributes of these connections (post, contact information, etc). The information mentioned is depicted in the visiting card of the person if a person holds a few posts (in one or different organizations), the information of all posts is depicted.

The bibliographic part consists of different databases: publications of employees of the institute, content of the journals published in SB RAS, own bibliographic database of the system "Web-resources of the mathematical content" etc. At the present time information from different data bases is represented independently, but we plan to create a common output system eliminating duplicates.

References

[1] Portal MathTree. www.mathtree.ru

[2] Portal Math-Net.RU. www.math-net.ru

[3] Portal MacTutor History of Mathematics. www-history.mcs.st-and.ac.uk/history/

[4] Zhizhimov O.L., Mazov N.A. Principles of Construction of the Distributed Information Systems on the Basis of the Protocol Z39.50. Novosibirsk: Publ. ICT SB RAS, 2004. 361 p.

[5] Valiev M.K., Kitaev E.L., Slepenkov M.I. Usage of the Service of Directories LDAP for Representation of a Metainformation in Global Computing Systems. www.keldysh.ru/metacomputing/ism99.html

[6] Ullman J.D. Principles of database system. Comp. Sci. Press, 1980.

[7] Maier D. The Theory of Relational Databases. Comp. Sci. Press, Rockville, MD, 1983.

[8] The concept of open systems // Materials to the Interbranch Program "Development and Application of Open Systems". www.informika.ru/windows/inftech/opensys/3/concept/os_1.html

[9] Microsoft.Net. www.microsoft.com/net/

[10] XForms 1.0. W3C Working Draft 16 February 2001. www.w3.org/TR/2001/WD-xforms-20010216

[11] Resource Description Framework (RDF) Model and Syntax Specification. W3C Recommendation 22 February 1999. www.w3.org/TR/1999/REC-rdf-syntax-19990222/

[12] Barakhnin V.B. Development of the thesaurus of data domain of "Mathematician" // Materials Intern. Conf. "Computational and Informational Technologies for Science, Engineering and Education". Pt 1. Ust-Kamenogorsk, Kazakhstan, Sept. 11–14, 2003. P. 111–115. (in Russian).

[13] Barakhnin V.B., Guskov A.E., Klimenko O.A. et al. Information system "Web-resources of mathematical Content" // Proc. of the Conf. of Young Sci. Devoted to M.A. Lavrentyev, Novosibirsk, 2004. Pt I. P. 23–27. (in Russian).

[14] Information System "Web-resources of the Mathematical Content". www.nsc.ru/win/mathpub/mathem_www.html

[15] Information System "Mathematicians of SB RAS". www.sbras.ru/sbras/math_soran/eng/

[16] Mathematics Subject Classification. www.ams.org/msc/

[17] Shokin Yu.I., Fedotov A.M., Klimenko O.A., Leonova Yu.V. Informative filling of the reference system of scientific association // Materials Intern. Conf. "Computational and Informational Technologies for Science, Engineering and Education". Pt 4. Alma-Ata, Kazakhstan, Oct. 6–10, 2004. P. 346–350. (in Russian).

Received for publication 19 August 2006