SOFTWARE TOOLS FOR NAVIGATION IN DOCUMENT DATABASES

Development of Information Navigation Service Based on Classification Schemes

Pavel Shapkin
MEPhI, Kashirskoe Shosse, 31, Moscow, Russia

Alexander Shapkin
VINITI, Usievicha, 20, Moscow, Russia

Keywords: Classification schemes, classification scheme mapping, information search, XML web services, Semantic Web.

Abstract: Internet allows accessing large document databases contained in different information centres across the world. Each database has its own search engine which is based on an index or classification scheme. Problems occur when a user tries to search different databases at once: different databases use different classification schemes. This article describes a classification scheme mapping service which is useful in integration of different databases in one search engine.

1 INTRODUCTION

Many information centers in the world are processing scientific information. As an example of these centers we can consider CAS, BIOSYS, Medline, ISI etc. In Russia the All-Russian Scientific and Technical Information Institute of Russian Academy of Sciences (VINITI) is such center. Each centre contains large amount of scientific information in form of document databases. Each centre creates an index of the documents based on a classification scheme, e. g. Universal Decimal Classification (UDC) or Mathematics Subject Classification (MSC). By means of this classification scheme users can search documents in database (Batty, 1998; Gilyarevsky, 1971). But if a search involves different databases simultaneously, problems occur: different information centers use different classification schemes.

Nevertheless problem of searching different databases at once becomes more and more actual because of expansion of Internet, which connects different information centers in one worldwide network (Clarke, 2000). In order to give users an opportunity to search different document databases at once we need a public resource which can provide a service for converting concepts from one classification scheme to another. It is important that user has not to be familiar with distinctions in indexing and subject representation in different information centers.

It is sometimes difficult to define concordances between different classification schemes. This work must be conducted by experts. That's why mapping service can be separated in form of an independent resource and used as a component of meta-search engine. This service has to be maintained by organization which keeps track on changes in classification schemes and uses experts to define concordances.

Using a mapping service between different classification schemes user has to know only one classification scheme. An attempt of building such a service is carried out in VINITI. VINITI uses different classification schemes during the processing of incoming document flow. Thus a large amount of knowledge about concordance of classifications is accumulated. Furthermore, VINITI is a member of UDC Consortium, and is responsible for maintaining Russian version of this classification.

Shapkin P. and Shapkin A. (2007).
SOFTWARE TOOLS FOR NAVIGATION IN DOCUMENT DATABASES - Development of Information Navigation Service Based on Classification Schemes.
In *Proceedings of the Third International Conference on Web Information Systems and Technologies - Web Interfaces and Applications*, pages 455-458
DOI: 10.5220/0001285804550458

2 STRUCTURE OF CLASSIFICATION SCHEME SYSTEM

System of Classification Schemes (SCS) includes classifications that are used in processing of incoming flow of heterogeneous scientific information. System allows maintaining a set of classifications which have complicated structure and also comparing them.

Atomic elements of classifications are called *rubrics*. Classification is a hierarchical structure of rubrics based on "parent-child" relation among rubrics. Each rubric has a code — unique identifier within the bounds of a classification.

The "parent-child" relation is main but not unique relation between rubrics. Often it is needed to represent more complicated interrelations of concepts which are beyond strict hierarchical scheme. That's why simple hierarchical model of classification scheme gets extended in SCS through introduction of *direct links* and other concepts. Furthermore SCS has means for *classification scheme mapping*. Detailed description of these concepts is given below.

2.1 Rubric Properties

Each rubric has some backbone properties and a set of optional properties which can vary according to the type of classification scheme.

Backbone properties are rubric code and code of parent rubric. Within the bounds of one classification scheme rubric codes are unique. Each rubric must have one parent. The only exceptions are root rubrics of classifications, which have no parents.

Other common property of rubrics is its title which can be given in different forms: full or short and on different languages.

Some classification schemes support descriptors (or keywords): each rubric can be concerned with a list of keywords. This approach simplifies search of rubrics.

Each classification scheme changes in time due to evolution of subject representation in concerned domain. Formally these changes mean addition, modification or removal of rubrics.

To maintain lifecycle of a rubric, special properties are introduced:

- date of creation;
- date of exclusion from classification scheme;
- current status, or lifecycle stage.

Besides the "parent-child" relation *direct links* between rubrics within a classification scheme are allowed. These links represent references like "see also", "reference from", "instead of" etc.

2.2 Classification Scheme Mapping

Comparison of classification schemes allows building links between rubrics contained in different classification schemes, in other words, mappings between classification schemes.

Examine two classification schemes *R* and *Q*. Each of them is a finite set of rubrics:

$R = \{r_1, r_2, \dots r_{[R]}\}$ where [R] is cardinality of *R*,

$Q = \{q_1, q_2, \dots q_{[Q]}\}$ where [Q] is cardinality of *Q*.

Mapping of rubric r_k (source rubric) from classification scheme *R* to classification scheme *Q* (target scheme) is a set of pairs (o_i, p_i) where o_i is an operator and p_i is nonempty set of rubrics from *Q*. Operators define meaning of relation, e. g. "includes" or "is equivalent to".

Consider an abstract example of mapping a rubric r_1 from *R* to *Q*, (operators are underlined):

r_1 includes q_1 and q_2
is included in q_6, q_7, q_8

Thus mapping of entire classification scheme *R* to classification scheme *Q* is defined as a set of mappings of all its rubrics to *Q*:

$$R \to Q = \{r_i \to Q \text{ where } i=1, 2, \dots [R]\}$$

As follows from the above, mapping has a direction — from source classification scheme to target classification scheme and mapping is one-to-many relation: it contains one source rubric and many target rubrics.

Initially mappings are created by experts. But expert work is expensive whereas system contains more than 20 classification schemes. Experts cannot create mappings between every pair of classifications. That's why some mappings are computed automatically from expert mappings, these are inverse and transitive mappings.

Consider an expert mapping *M* of classification scheme *R* to classification scheme *Q*. Inverse mapping for *M* is mapping of *Q* to *R*, computed from *M*. That is if rubric r_1 is mapped to q_1 with operator op_1 in *M*, then is inverse mapping q_1 is mapped to r_1 with inverse operator. Each operator has corresponding inverse operator, e. g. “is included in” is inverse of “includes”, etc.

Transitive mapping is a union of a chain of expert mappings. Suppose there exist expert or inverse mappings between schemes *R* and *Q*, and between *Q* and *P*. Then a transitive mapping of *R* to *P* can be computed from these two mappings. During computation of transitive mapping compositions of mapping operators are used. For example, if rubric r_1 is equivalent to q_1, and q_1 includes p_1, then r_1 includes p_1. Thus, “includes” is composition of “is equivalent to” and “includes”.

Introduction of transitive and inverse mapping gives an opportunity to construct mapping between virtually any pair of classification schemes with minimal cost: experts need to build only the basic mappings, all other mappings will be evaluated “on the fly”.

2.3 Quantitative Characteristics of Classification Scheme System

At present system contains 25 classification schemes, such as Universal Decimal Classification (UDC), International Patent Classification (IPC), AMS Mathematics Subject Classification (MSC), Library and Information Science Abstracts (LISA), VINITI Classification Index and others. Rubrics count in a scheme varies from 200 to 67000. Total number of rubrics contained in system is about 1 million.

Main list of descriptors contains 436000 keywords and phrases, number of keywords bound to a rubric varies from 1 to 150.

Inner links between rubrics are used in 11 classification schemes; mean part of rubrics connected with direct links amounts to 14% of total rubric count in classification scheme.

There are 19 expert mappings between separate classification schemes.

3 WEB SERVICE IMPLEMENTATION

SCS has a web-interface, which is an ASP.NET-based application (Payne, 2001); data is represented in XML format (Bean, 2003). The main part the application is a set of two web-services:

- metadata service;
- mapping service.

By means of the metadata service users can receive information about classification schemes which are available in the system. One can get a list of available schemes, lists of their properties and information on available mappings.

Mapping service allows users to get mapping of a rubric from one classification scheme to another.

Both services allow access using SOAP, HTTP GET or HTTP POST protocols.

Because implementation is based on an XML web-service, it can be accessed from any type of applications, e. g. from client application written on any programming language, from a web-site or even from AJAX-based web page. XML data can be transformed using XSL (Holzner, 2001).

Along with web-service a simple HTML interface was created. It uses AJAX to utilize metadata service in order to obtain information about available classification schemes and their mappings. User can choose source and target classification scheme from a list, enter code of source rubric and then receive the list of rubrics on which source rubric is mapped. An experimental test page of the service is located at “http://solar.viniti.ru:8080/MapService/mapform.aspx”.

4 USING SERVICE IN INFORMATION PORTAL

Let us examine an example of utilizing classification scheme mapping service within an information portal. Consider an internet resource containing a large database of publications, e. g. of dissertations. Each dissertation contains a Universal Decimal Classification (UDC) index, which determines the area to which this dissertation belongs. Suppose that portal implements a new service that enables users to get a list of patents whose area is close to the area of chosen dissertation. Patents can be obtained from a publicly available service like esp@cenet. The problem is that patents are indexed with International Patent Classification (IPC). That’s where the classification scheme mapping service appears. To get a list of patents it is needed first to send request containing the UDC index of chosen dissertation to mapping service. As a response portal system will receive a list of IPC indexes related to this UDC index. Then portal system can use this list in order to retrieve required patents from patent database. Interaction diagram for this example is shown on fig. 1

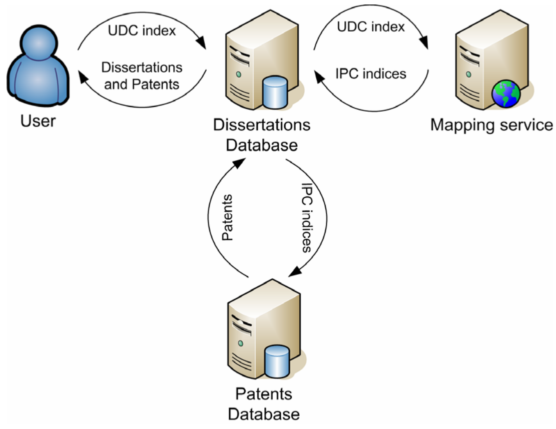

Figure 1: Classification scheme mapping service utilization example.

5 SEMANTIC WEB

The main idea of Semantic Web is to make information contained in web resources suitable not only for use by people but also for processing by machines. Instead of using HTML, which describes only the representation of data, Semantic Web languages allow to describe semantics of data explicitly.

One of Semantic Web principles is decentralization of data and centralization of metadata. It means that public metadata resources are needed. Classification schemes are perfect examples of metadata, thus service for accessing classification schemes and their mappings can act as a metadata provider. In order to integrate classification scheme service in Semantic Web it may be necessary to bring XML data representation format in correspondence with standards of Semantic Web. It implies using languages like RDF (Resource Description Framework, see W3C RDF Primer) and OWL (Ontology Web Language, see W3C OWL Guide).

The aim of RDF is to standardize format of describing metadata used in web resources. Main construction of RDF is triplet “object-attribute-value”. It can be written as A(O, V) which means “object O has an attribute A with value V”. Attributes are often called *properties* or *relations*, and objects are also called *entities*. Each element in the triplet can be specified with its Uniform Resource Identifier (URI) — global unique identifier. Triplets can be nested and thus form a graph.

An example of graph describing a rubric from VINITI Classification Index is shown on fig. 2. Entities are represented by ovals, values — by boxes and relations are represented by arrows. Identifiers are shown within the objects; types of objects are shown near them. Entities’ and relations’ URIs are shown. Use of URIs guarantees that, e. g. `Child` relation will be recognized as the parent-child relation used in VINITI Classification Index.

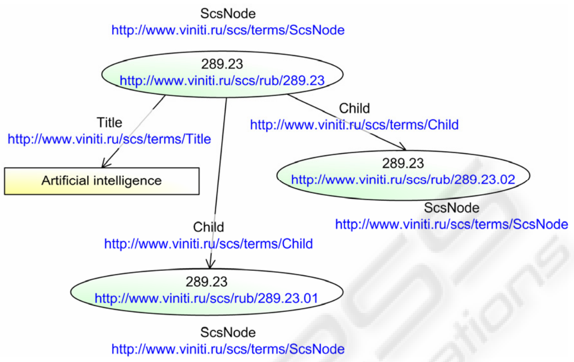
ScsNode
http://www.viniti.ru/scs/terms/ScsNode
289.23
http://www.viniti.ru/scs/rub/289.23
Title
http://www.viniti.ru/scs/terms/Title
Child
http://www.viniti.ru/scs/terms/Child
Artificial intelligence
289.23
http://www.viniti.ru/scs/rub/289.23.02
ScsNode
http://www.viniti.ru/scs/terms/ScsNode
Child
http://www.viniti.ru/scs/terms/Child
289.23
http://www.viniti.ru/scs/rub/289.23.01
ScsNode
http://www.viniti.ru/scs/terms/ScsNode

Figure 2: RDF representation graph for VINITI Classification Index rubric.

6 CONCLUSION AND FURTHER WORK

Classification scheme system can act as a source of information about different classification schemes for scientific information. Ability of mapping different classification schemes allows to use this system for purposes of integration document databases of different information centers.

REFERENCES

Payne, C., 2001. *Teach Yourself ASP.NET in 21 Days*, Sams.

Bean, J., 2003. *XML for Data Architects: Designing for Reuse and Integration*, Morgan Kaufmann.

Holzner, S., 2001. *Inside XSLT*, Que.

RDF Primer. W3C Recommendation 10 February 2004, from http://www.w3.org/TR/2004/REC-rdf-primer-20040210/

OWL Web Ontology Language Guide. W3C Recommendation 10 February 2004, from http://www.w3.org/TR/2004/REC-owl-guide-20040210/

Batty, D. Controlled vocabulary and thesauri in support of online information access. *D-Lib Magazine*, November, 1998

Clarke S.J. Search Engines for the World Wide Web: An Evaluation of Recent Developments. *Journal of Internet Cataloging*, sol.20, №3/4, 2000, 81-93.

Gilyarevsky R.S. UDC and its role in the development of the information retrieval languages. *Paper synopses*. Herceg Novi (Jugosl.), 1971. P. 6.