Thesaurus Implementation in Integrated System of Information Resources (ISIR)

Nguyen Manh Hung

Computing Center, Russian Academy of Sciences, ul. Vavilova 40, Moscow, 119991 Russia
e-mail: nmhungru@yahoo.com

Received February 17, 2004

1. PROBLEM STATEMENT

Thesauri are used, first of all, for classifying and searching for information resources. Each resource may be associated with one or more concepts described with the use of certain terms in a thesaurus, and the user can find the desired concepts in a given subject domain by means of the thesaurus, as well as all terms characterizing them. Thus, retrieval is extended through the use of thesaurus links (the set of words in a query is extended by synonymous, more general, or more particular terms). The navigation through the thesaurus links helps to better formulate the query itself. Thus, thesauri in information systems can be used for solving various tasks, such as classifying information, extending queries [1], determining key words for a text, automated annotating texts [2], and the like.

Our objectives are to study thesaurus structure, to try to create a general universal schema of thesauri, and to implement this schema in an information system. The thesaurus implementation should be universal in order to make it possible to work with any thesauri satisfying the existing standards. This means that the thesaurus schema is to be universal as well; i.e., it should make it possible to represent the structure of any thesaurus. In order that the system could exchange information with other systems, the thesaurus implementation must meet the Semantic Web standards.

2. USE OF THESAURUS IN INFORMATION DESCRIPTION

In order to describe some subject domain, a certain set of terms is used, which describe certain concepts from this subject domain. A *thesaurus* is set of terms describing a given subject domain together with semantic relations (links) between them. The relations in a thesaurus always indicate the existence of semantic links between the terms.

The basic relation (link) between the terms in a thesaurus is the relation between broader (more expressive) and narrower (more specialized) terms. The three following kinds of this relation are often distinguished [3]:

- One term denotes a concept that is a part of a concept denoted by another term (for example, “science” and “mathematics,” or “mathematics” and “number theory”).
- One term denotes a particular case of a concept denoted by another term (for example, “birds” and “parrots”).
- One term denotes an element of a class denoted by another term (for example, “mountain districts” and “Caucasus”).

This relation on a set of terms is a partial order relation; i.e., a set of terms with such relations forms an acyclic graph, or a polyhierarchical structure.

There also exist other relations between the terms. For example, one concept can be denoted by several synonymous terms. Then, among these terms, one (for each thesaurus language) most preferable (most appropriate) term called a descriptor (preferred term) [4], which better describes or characterizes this concept, is separated. The other terms are less preferable (less appropriate) terms called ascriptors (non-preferred term) [4]. There may exist other relations between an ascriptor and descriptor, such as, for example, “use a combination” or “use alternatively” [4].

In addition to the above-mentioned relations, there may exist other, associative links between terms if the meanings of the concepts denoted by these terms are related to each other somehow, except for the hierarchical relations described above.

In multilingual thesauri, there also exist equivalence relations between terms in different languages [5, 6]. It is commonly accepted to separate complete (strong) equivalence and several kinds of partial (weak) equivalence of terms in different languages.

A thesaurus often contains comments on terms that give the user some additional information about the meaning of the term and how to use it.

There exist various standards (which differ in their significance levels and degrees of detail) on the format of thesaurus representations. These standards represent a thesaurus as a set of objects of several types with relations of several types between the objects. Some standards (for example, ANSI/NISO Z39.19-1993, GOST 7.25-2001) also regulate a linearized (text) format of
the thesaurus representation that can be understood by both machines and humans.

Specific features of these standards are discussed in detail in [7].

3. REQUIREMENTS ON THE IMPLEMENTATION OF THESAURI IN INFORMATION SYSTEMS

A particular implementation must satisfy the following requirements:

1. Any existing thesauri and, in particular, any classifiers possessing a thesaurus structure in accordance with the GOST 7.25-2001, GOST 7.24-90, ISO 2788, and ISO 5964 standards can be stored in the implementation. In particular, it should be capable of working with multilingual thesauri.

2. The implementation should make it possible to index resources by terms of the thesaurus and to classify them in terms of the concepts of the thesaurus-classifier. In addition, both kinds of the thesauri should be worked with in a uniform way.

3. It should allow one to navigate the thesaurus and search for resources indexed or classified by the thesaurus. In other words, the implementation must ensure effective execution of all required queries, namely,

- to obtain all concepts related to a given concept through links of specified kinds (for links, in accordance with the GOST standards);
- to obtain the uppermost concepts in the hierarchy of concepts containing the given one;
- to obtain all terms related to a given concept through links of specified kinds (for links, in accordance with the GOST standards);
- to obtain all concepts related to a given term through links of specified kinds (for links, in accordance with the GOST standards);
- to obtain all terms containing a given word (or a key word);
- to obtain a complete hierarchy of concepts of the thesaurus, i.e., to show terms in the hierarchy.

4. It should make it possible to extend and redefine user’s retrievals to the system using links between concepts (terms) of the thesaurus.

5. It should be possible to load and unload thesaurus data in the RDF/XML formats for exchanging terminological data with other systems.

6. The implementation is to be extensible, i.e., to allow refinement if certain links are required.

4. THESAURUS MODEL

4.1. Thesaurus Implementation Platform

A thesaurus model, including a model that takes into account all requirements listed above, can be created on almost any platform for ontology representation. In particular, there exist thesaurus models based on Topic Maps [8], RDFS [9, 0], and DAML [11].

However, to ensure better correspondence of the thesaurus implementation to the concepts of the Semantic Web project, the following requirements are imposed on the implementation:

1. Syntactic and semantic interoperability. Any application working in accordance with the Semantic Web requirements must be able to work with the thesaurus without preliminary adjustment of the formats.

2. Thesaurus extensibility. If required, any application must have an opportunity to add its elements to an open thesaurus and use it in such an extended form for its own needs.

3. Model extensibility. A data schema must admit extensions and circumstantiation; i.e., any application must have an opportunity to add new types of resources and links to the model and, in particular, redefine the existing ones if this is required for the description of a nonstandard thesaurus. At the same time, the applications that do not know about such an extension must have an opportunity to work with this thesaurus in the framework of the previous model having access to only that part of thesaurus data that agrees with that model.

These requirements impose also certain restrictions on the thesaurus implementation platforms. For example, the Topic Maps platform [8] in the XTM 1.0 format, on the whole, satisfies the first and second requirements but does not satisfy the third requirement, since the possibilities of inheriting class properties and adding comments to links are lacking.

From the standpoint of these requirements, the RDFS platform, along with its extensions (for example, DAML+OIL), is most appropriate. The RDFS platform is also accepted as the basic one for the ontology description in Semantic Web.

4.2. Thesaurus RDF-Schema

This data schema is based on the RDFS platform. Our approach to describing thesauri relies on the work [7]. When describing thesauri, we rely on the GOST standards, because the GOST standards are, in fact, extensions of the ISO and ANSI/NISO standards.

In accordance with GOST 7.25-2001, a thesaurus has terms of two kinds: descriptors and ascriptors. Relations between descriptors include hierarchical, associative, and equivalent relations. A descriptor may have the following relations with ascriptors: synonymous relations, “compare alternatives,” and “compare combinations.” An ascriptor may have the following relations with descriptors: “look,” “use alternatively,” and “use a combination.”

According to GOST 7.27-2001, there are two basic classes of objects in the schema: ThesaurusConcept, which is a concept expressed by a descriptor, and ThesaurusTerm, which is a concept expressed by an ascriptor. Then, relations between descriptors in GOST 7.25-