Intelligent Information Retrieval Based on Interconnected Concepts and Classes of Retrieval Domains

Kuldar Taveter, VTT Information Technology, kuldar.taveter@vtt.fi

Keywords: information seeking behaviour, visualisation, metaknowledge

1. Introduction

The structure of information in any information source represents concepts and inter-concept relationships of the domain in question that the creator of the information source had in his mind. On the other hand, people often express their information needs in terms of concepts that information is needed about. The *conceptualization* of the world embodied in some information source may be in the form of a database schema, or it may also be some classification that the information in the source is based on. One of the most important problems that has to be solved in semantical information retrieval from heterogeneous sources is to reconcile different conceptualizations of the world represented by different information sources [1]. A part of this problem is that different information sources make use of different classifications of the same objects of the world. One way to tackle this is to distinguish the ontological aspects of objects (i. e. the conditions of their being, structure, integrity, identity) from the taxonomic aspects (i. e. the conditions for seeing them as members of one or another particular class) [3]. This is the approach we have chosen in SARI, which is an agent-based system of semantical information retrieval that is being jointly worked out by VTT Information Technology, Tampere University of Technology, and Tampere University.

2. Ontology vs. taxonomy

According to Guarino et al [2] *ontology* can be understood as an intensional semantic structure which encodes the implicit rules constraining the structure of a piece of reality. Ontologies are thus aimed at answering the question "What kinds of objects exist in one or another domain of the real world and how are they interrelated?". Ontologies can be made explicit by forming a logical theory which gives an explicit and partial account of the above-mentioned intensional semantic structure. Such logical theory contains concepts, their definitions, and relationships between them like e. g. subsumption (inheritance) and aggregation. Ontologies contain concepts of two kinds: *types* and *roles*. The basic difference between them is that the former are semantically rigid, i. e. their instances are necessarily such that they always belong to them, while this is not the case for the latter. For example, a plant will be a plant during its whole lifetime, while a student can cease to be a student and still remain the same individual [9].

In [3] it is claimed that ontologies should be separated from taxonomies because since there exist several different ways to classify the same objects in concurrent taxonomies, the objects must be independent from these taxonomies. In other words: objects that belong to a certain concept can be classified in very different ways depending on the *viewpoint*. For example, genes in biology can be classified differently from functional, chemical, and evolutionary viewpoints. We are now trying to bring this claim to a firmer ontological basis by further claiming that each of the concurrent taxonomies classifies objects of some concept according to a particular *role* that is subsumed by this concept, and represented by the taxonomy's root class. For example, the taxonomies under the classes "Commodity" and "Product" in Figure 1 classify certain man-made objects (artifacts) according to the roles that they respectively play in the domains of foreign trade and manufacturing. Both of the mentioned classes are subsumed by the concept "Artifact" of the ontology by the Role-Of relationship (v. Figure 1).

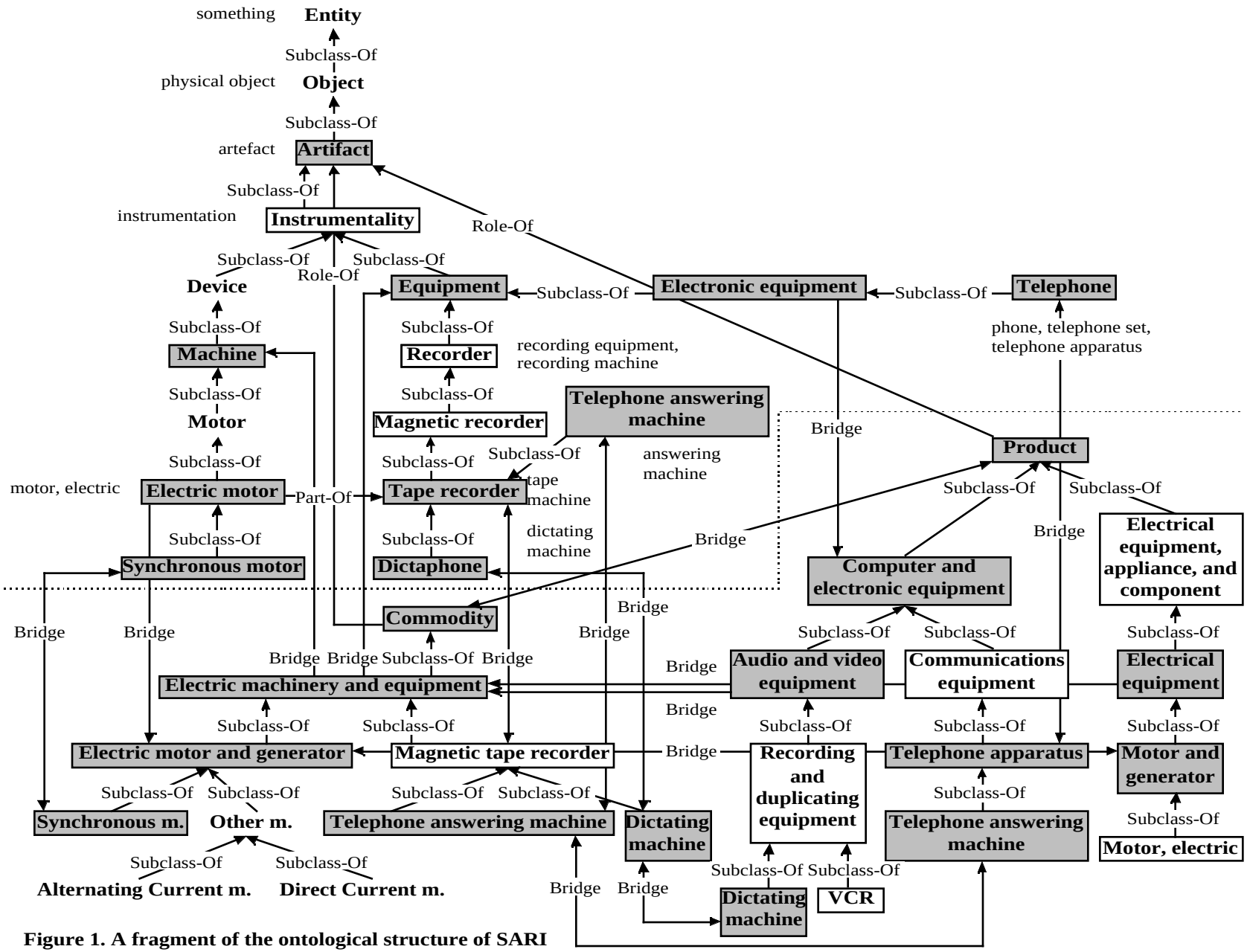

Figure 1. A fragment of the ontological structure of SARI

3. Viewpoints and bridges in information retrieval

Since a hierarchy of classes under which the instances of a concept (objects) can be classified is determined by a particular viewpoint, each role corresponds to a viewpoint. Consequently, we can say that the part of Figure 1 below the dotted line depicts the classifications (taxonomies) of certain artifacts under the *foreign trade viewpoint* (left) and *manufacturing viewpoint* (right).

In order to represent links between the class structures of different taxonomies, we make use of the notion of bridge. A *bridge* between two classes under different viewpoints means that an object which is a member of the source class under one viewpoint is also a member of the destination class under the other viewpoint [4]. In Figure 1 there is a bridge between the classes "Commodity" and "Product" which are respectively the root classes under the foreign trade and manufacturing viewpoints. While in [3] and [4] bridges and viewpoints are used for creating classifications of real world objects, we use them to describe already existing classifications that are conceptualized by information sources. For example, the classification under the class "Commodity" in Figure 1 is an excerpt from the standardized CN hierarchy of commodity types [5] which is used by the statistical database Ultika of Finnish foreign trade, and the classification under the class "Product" in the same figure is a subset of the NAICS classification of industry [6] used in North America. Bridges can be divided into one-way bridges and two-way bridges. We say that there is a *one-way bridge* between two classes under different viewpoints when all possible member objects of the source class also belong to the destination class. There is a *two-way bridge* between two classes when this is true in both ways, i. e. when the sets of possible extensions[1] of the two classes are equal. For example, there is a one-way bridge between the concepts "Electrical equipment" and "Electric machinery and equipment" under the viewpoints of manufacturing and foreign trade, respectively, because all possible instances of the first class also belong to the second class. On the other hand, since both the class "Electric motor and generator" under the foreign trade viewpoint and the class "Motor and generator" under the manufacturing viewpoint represent the set of *electric* motors and generators, there is a two-way bridge between them. Please note also that there is *no* bridge between the classes "Magnetic tape recorder" and "Recording and duplicating equipment" under the respective viewpoints of foreign trade and manufacturing because:

- while the first class involves telephone answering machines, the second class doesn't;
- while the second class involves video cassette recorders (VCRs), the first class doesn't.

The concepts that are involved in any bridge relationship are marked grey in the figure. The only restriction on forming bridges is that a bridge should not bring about inheritance contradictions between the class hierarchies of different viewpoints, i. e. bridges should not cross each other.

Let us now consider a task where the user wants to query Ultika and some database using NAICS in parallel which is quite a realistic situation. The possibility to browse in parallel the conceptual structures of both databases and have the needed queries generated makes the process of information retrieval cognitively natural and easy for the user. While browsing the user is able to switch between viewpoints at different locations of the databases' conceptual structures using bridges.

We also make use of the bridge relationships to link the concepts of the ontology to the corresponding classes of the taxonomies. Incorporating parallel conceptual structures into the *ontological structure* consisting of the ontology and its taxonomies considerably speeds up parallel browsing. For example, by using bridges, the user can immediately proceed from the browsing of the ontology's concept "Electric motor" to the browsing of the corresponding classes "Electric motor and generator" and "Motor and generator" of

[1] The extension of a class is any set of its individuals (objects, occurrences, instances).

the taxonomies of commodities and products, respectively. In our example an excerpt from the ontology that both of the taxonomies refine is depicted in the part of Figure 1 above the dotted line. In order to retain legibility of Figure 1, only some bridges between the concepts of the ontology and the classes of the taxonomy of products are represented there.

4. The use of terms and synonyms

The ontology presently used by us is based on WordNet [10]. The ontology represents concepts and relationships between them. In the figure subsumption (Subclass-Of) and aggregation (Part-Of) relationships are depicted. Each concept of the ontology is represented by a natural language expression called *term* which is the text in the corresponding rectangle, and possibly by one or more *synonyms* which are given at the side of the rectangle. The term is the most typical or obvious of the synonyms. The classes of the taxonomies in Figure 1 are also represented by their terms. Synonyms play an important role in concept-based query formulation and expansion [7], and in replying to free-form queries [8]. For example, a user of SARI who is looking for information about dictaphones can start off by entering or having generated with the help of the ontological structure the query "dictaphone" which is thereafter expanded by the synonym of the term "Dictaphone" to the query "'dictaphone' OR 'dictating machine'". The synonym "dictating machine" matches with the terms of the corresponding classes of both taxonomies. In the same manner, a user who enters the query "phone" will eventually be taken to the class "Telephone apparatus" under the manufacturing viewpoint because both "phone" and "telephone apparatus" are synonyms for the term "Telephone" of the ontology. In SARI all queries are implicitly processed as case-insensitive.

5. Conclusions and future work

The main contributions of our work are the following:

- further elaboration and clarification of the distinction and connections between ontology and taxonomy by utilizing the notion of role;
- the use of the notions of viewpoint and bridge in information retrieval;
- the coinage of the notions of one-way bridge and two-way bridge to be used for connecting the conceptual structures of different information sources;
- the use of bridges for connecting the concepts of an ontology to the classes of its taxonomies, resulting in the ontological structure.

The ontological structure consisting of the ontology and the taxonomies refining its concepts can be applied to:

- query expansion and generation for structured (e. g. relational, hierarchical, OO) databases, and for browsing their conceptual structures;
- query expansion for textual databases;
- query expansion for WWW.

The most important problem that remains to be solved in our future work is how to make the formation of bridges semiautomatic. In principle this can be done by successive comparing each of the ontology's terms and its expanding synonyms with the terms of its taxonomies. The main obstacles lie in the computational complexity of such a task on one hand, and in the abundance of different grammatical forms in which the terms and synonyms can be expressed. This is especially true for agglutinative languages like e. g. Finnish and Estonian. Our future work will also include the formalization of the distinction between ontology and its taxonomies using roles.

The results of our research work will be implemented in the third pilot of the SARI system.

6. References

1. F. Saltor, E. Rodríguez. On Intelligent Access to Heterogeneous Information. Proceedings of the 4th KBRD Workshop. Athens, Greece, August 1997.
2. N. Guarino, P. Giaretta. Ontologies and Knowledge Bases: Towards a Terminological Clarification. In: N. J. I. Mars (ed.), Towards Very Large Knowledge Bases, IOS Press 1995, pp. 25-32.
3. J. Euzenat. On a purely taxonomic and descriptive meaning for classes. Proceedings of the IJCAI workshop on "Object-based representation systems". Chambéry, France, 1993.
4. J. Euzenat. Brief overview of T-tree: the Tropes Taxonomy building Tool. In: Philip Smith, Clare Beghtol, Raya Fidel, Barbara Kwasnik (eds.), Advances in Classification Research 4, Learning Information, Medford (NJ US), 1994.
5. Combined Nomenclature (CN). Commission Regulation (EC) No 3115/94 of 20 December 1994. Official Journal of the European Communities L 345, 31 December 1994.
6. See http://www.census.gov/epcd/www/naics.html
7. J. Kekäläinen, K. Järvelin. The impact of query structure and query expansion on retrieval performance. Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia, August 24 - 28, 1998.
8. J. Kristensen. Expanding end-users query statements for free text searching with a search-aid thesaurus. Information Processing & Management, 29(6): 733_744.
9. N. Guarino, C. Masolo, G. Vetere. OntoSeek: Using Large Linguistic Ontologies for Gathering Information Resources from the Web. LADSEB-CNR Technical Report 01/98, March 1998. Submitted for publication.
10. See http://www.cogsci.princeton.edu/~wn/.